

Saaty

Lectures on  
Modern Mathematics

Volume  
II

---

# Lectures on Modern Mathematics

*Volume II*

*Edited by T. L. SAATY*

---

510

LEC

Wiley

*Saaty*

Lectures on  
Modern Mathematics

Volume  
III

510

LEC

*Wiley*

*Saaty*

Lectures on  
Modern Mathematics

Volume  
II

510

LEC

*Wiley*

# **Lectures on Modern Mathematics**

Γ Δ Θ Λ Ε Π Σ Φ Ψ Ω

# Lectures on Modern Mathematics

*edited by T. L. Saaty, Office of Naval Research  
and Arms Control and Disarmament Agency*

**volume II**

*John Wiley & Sons, Inc., New York • London • Sydney*

*Copyright © 1964 by John Wiley & Sons, Inc.*

All Rights Reserved. This book or any part thereof must not be reproduced in any form without the written permission of the publisher, except for any purpose of the United States Government.

*Library of Congress Catalog Card Number: 63-20639*

*Printed in the United States of America*

|               |   |         |
|---------------|---|---------|
| ACCESSION No. |   | 64817   |
| CLASS No.     |   | 510 LEC |
| 23 FEB 1976   |   |         |
| CS            | ✓ |         |

# Preface

"There is no inquiry which is not finally reducible to a question of number" wrote August Comte. However, the reader who has examined the first volume of this series and who leafs through the pages of this volume will find that, with some exceptions, the qualitative inquiry of modern mathematics itself is not always strictly reducible to a quantitative question of number. A greater effort is directed in modern mathematics toward structuring and characterizing systems than toward examining their quantitative implications. Either we must interpret Comte's "question of number" as the logical ordering and description of the inquiry in a rigorous manner rather than its measurement, or we must conclude that perhaps in his day he could not comprehend what mathematics was to be about. Are the other sciences more susceptible to quantitative inquiry?

There is no more eloquent and symbolic monument than one constructed in the spirit of an age. We may hope that the lectures manifest this spirit and provide building stones for the monument of modern mathematics.

The many unsolved problems and new challenges should make us more imaginative than Comte in delineating its nature for the future.

In the preface to Volume I, I made the following observations, which have overall significance for the series, and should perhaps be repeated.

In our time the growth of mathematical literature, like that of all the sciences, has burgeoned beyond the point where an individual finds its content accessible. It has been shown that the number of scientists and the number of published scientific papers have multiplied by ten for every doubling of the general population, and hence that the next decade or two will see the doubling of the volume of all the existing literature produced until now. The man who is not a specialist in a given area of research will find that the sheer volume of published work makes him a stranger, and the proverbial esotericism of mathematics has by now limited the possibility of universalism to the very few.

For the future we are promised the applied marvels of data processing and information storage and retrieval. Nonetheless, it requires more than mechanical assimilation to unify and sift for significance the published knowledge of even a relatively small sub-field of mathematics. We are still obliged (and shall be for a long visible future) to look to the mature specialist himself for such evaluation, summarization, and interpretation as a partial answer to the "paper explosion."

This is the second volume of the series of lectures jointly sponsored by the George Washington University and the Office of Naval Research, which by intelligent summaries further extends the horizons of accessible modern mathematics.

The lectures, the first six of which appeared in the first volume, were begun in the fall of 1962 at the Lisner Auditorium of the George Washington University and continued at monthly intervals. This volume contains six more lectures, three given in the spring and three in the fall of 1963. A third volume for the remaining six lectures of the academic year 1963-1964 will follow.

Our intention, as stated in the preface to Volume I,

... was to invite each of the eminent men represented here to delineate a substantial research area, to describe it broadly and comprehensively for an audience of mathematicians who are not specialists in that area, and to contribute to this description his individual evaluation, of the esthetic and practical aspects of the field, its position in mathematical development as a whole, and its future, as that might be implied in the conjectural exposition of its unsolved problems. The speakers have responded to a difficult

challenge, compressing the enormous material at their command into a necessarily limited space, preserving the original spirit of the project in their informal approach, refraining from any deep and intricate excursions which might intimidate the tyro, and at the same time giving in each case the personal flavor of their own involvement in mathematical research.

The outstanding success of the lecture series makes us confident that this book will be of interest to all mathematicians desiring to keep abreast of the major achievements in various fields, as well as to those in the general scientific public wanting to have a flavor of the rapid and sophisticated development of the "Queen of the Sciences." To the graduate student in mathematics embarking on his research career it should be both useful and encouraging, providing glimpses of large areas to which he may not have been formally introduced and challenging him with unsolved problems.

Our thanks go again to Professor David Nelson for his continued and untiring efforts as the executive of this project.

T. L. SAATY  
EDITOR

*Washington, D. C.*  
*June, 1964*

# List of Lectures

## VOLUME I (Published 1963)

- P. R. Halmos, *A Glimpse into Hilbert Space*  
Laurent Schwartz, *Some Applications of the Theory of Distributions*  
A. S. Householder, *Numerical Analysis*  
Samuel Eilenberg, *Algebraic Topology*  
Irving Kaplansky, *Lie Algebras*  
Richard Brauer, *Representations of Finite Groups*

## VOLUME II

- L. Nirenberg, *Partial Differential Equations with Applications in the Geometry*  
Marshall Hall, Jr., *Generators and Relations in Groups—The Burnside Problem*  
R. H. Bing, *Some Aspects of the Topology of 3-Manifolds Related to the Poincaré Conjecture*  
Lars Gårding, *Partial Differential Equations: Problems and Uniformization in Cauchy's Problem*  
Lars V. Ahlfors, *Quasiconformal Mappings and Their Applications*  
John Milnor, *Differential Topology*

## VOLUME III (To be published)

- Georg Kreisel, *Mathematical Logic*  
M. M. Loève, *Stochastic Processes*  
Paul Erdős, *Number Theory*  
Einar Hille, *Classical Analysis*  
H. S. M. Coxeter, *Geometry*  
Joseph Kampé de Fériet, *Random Solutions of Differential Equations*

# Contents

- 1 Partial Differential Equations with Applications in Geometry, 1**  
*L. Nirenberg*
- 2 Generators and Relations in Groups—The Burnside Problem, 42**  
*Marshall Hall, Jr.*
- 3 Some Aspects of the Topology of 3-Manifolds Related to the Poincaré Conjecture, 93**  
*R. H. Bing*
- 4 Partial Differential Equations: Problems and Uniformization in Cauchy's Problem, 129**  
*Lars Gårding*
- 5 Quasiconformal Mappings and Their Applications, 151**  
*L. Ahlfors*
- 6 Differential Topology, 165**  
*J. Milnor*



# 1

## Partial Differential Equations with Applications in Geometry\*

*L. Nirenberg*

### *Lecture I General Linear Partial Differential Equations*

We shall present some of the recent developments in the general theory of partial differential equations. A survey of the extensive work of the last ten years on equations with constant coefficients has been given by Professor L. Schwartz in his lectures, and Professor L. Gårding has described current research on the initial value problem for hyperbolic equations. Thus we shall not treat these topics. Furthermore we shall confine ourselves to linear equations since the author has recently published an expository lecture on nonlinear equations [45]. Because of lack of space we can consider only a few topics.

In this lecture we consider mainly equations of arbitrary type, and treat questions of existence of solutions, locally and in the large, uniqueness for the initial value problem, smoothness of the solu-

\* This chapter represents results obtained at the Courant Institute of Mathematical Sciences, New York University, under the sponsorship of the Office of Naval Research, Contract No. Nonr-285(46). Reproduction in whole or in part is permitted for any purpose of the United States Government.

tions, and describe some of the techniques involved. In the second lecture we consider elliptic differential equations as they enter in several problems recently studied in differential geometry in the large. A number of the topics in this lecture are treated in Hörmander's book [25] which also presents many results on equations with constant coefficients. The following books or lecture notes also contain much of the recent work on such equations: Treves [53], Gelfand, Silov [19], Friedman [15].

We consider functions of  $n$  variables  $x = (x^1, \dots, x^n)$  and use the standard notation for differentiation:  $D_j = (1/i) \partial/\partial x^j$ ,  $D = (D_1, \dots, D_n)$ ; for a multi-index  $\alpha = (\alpha_1, \dots, \alpha_n)$ ,  $D^\alpha = D_1^{\alpha_1} \dots D_n^{\alpha_n}$  is an operator of order  $|\alpha| = \sum \alpha_j$ . We also set  $\alpha! = \alpha_1! \alpha_2! \dots \alpha_n!$  and  $\xi^\alpha = \xi_1^{\alpha_1} \dots \xi_n^{\alpha_n}$  for a vector  $\xi = (\xi_1, \dots, \xi_n)$ . A linear differential operator of order  $m$  is a polynomial in  $D$  with coefficients depending on  $x$ ,  $P(x, D)$ . If  $P'$  is the part consisting of terms of exactly order  $m$ , the principal part, then  $P'(x, \xi)$  is called the characteristic polynomial.

We shall use  $|u|_k$  to denote the  $C^k$  norm: supremum of derivatives of  $u$  up to order  $k$  in the domain. We shall also use the spaces  $H_s$  with norms

$$\|u\|_s^2 = \int \sum_{|\alpha| \leq s} |D^\alpha u|^2 dx$$

If the domain is the whole space this norm may also be expressed in terms of the Fourier transform  $\bar{u}(\xi) = \int \exp(-ix \cdot \xi) u(x) dx$  by

$$\|u\|_s^2 = \int (1 + |\xi|^2)^s |\bar{u}|^2 d\xi$$

This form enables us to define the norm for all real  $s$ .

A function defined in  $\mathfrak{D}$  will be said to belong to  $C_0^\infty(\mathfrak{D})$  if it is  $C^\infty$  and its support is a compact subset of  $\mathfrak{D}$ .

## 1. AN EXAMPLE WITH CONSTANT COEFFICIENTS

At the heart of the study of differential equations are inequalities related to the differential operators. We shall begin with a simple but instructive illustration of their use. Consider the well-known inequality of Hörmander [20] for operators  $P(D)$  with constant coefficients. We denote by  $P^{(\alpha)}(\xi)$  the derivative  $(\partial/\partial \xi_1)^{\alpha_1} \dots (\partial/\partial \xi_n)^{\alpha_n} P(\xi)$  of the polynomial  $P(\xi)$ . The inequality asserts that

there is a constant  $C$  depending only on  $n$  and  $m$  (the order of  $P$ ) such that for all  $C^\infty$  functions  $v$  with support in a sphere of radius  $r$  the inequality

$$\sum_{|\alpha| \geq 1} \|P^{(\alpha)}v\|^2 \leq Cr^2\|Pv\|^2(1 + r^{2m-2})$$

holds where  $\| \cdot \|$  denotes the  $L_2$  norm. Simple proofs may be found in [25] or [53]. Thus in particular, with a constant  $c$ ,

$$(1.1) \quad \|v\|^2 \leq cr^2\|Pv\|^2$$

With the aid of the Hahn-Banach theorem, this inequality easily implies the existence of  $L_2$  solutions as distributions of  $Pv = f$  for every  $f$  in  $C_0^\infty$ , that is, in  $C^\infty$  with compact support. By a slight modification it may also be used to give information in other problems, such as uniqueness and regularity in the Cauchy problem. If we set, namely,  $v = e^{\eta \cdot x}u$  with  $\eta = (\eta_1, \dots, \eta_n)$  a real vector,  $\eta \cdot x = \sum \eta_j x^j$ , so that  $e^{\eta \cdot x}Pu = P(D + i\eta)v$ , we find on applying Hörmander's inequality to the operator  $P(D + i\eta)$ ,

$$(1.2) \quad \sum_{|\alpha| \geq 1} \|e^{\eta \cdot x}P^{(\alpha)}u\|^2 \leq cr^2\|e^{\eta \cdot x}Pu\|^2$$

This yields immediately the fact that if  $u$  is a solution of  $Pu = f$  with  $u$  and  $f$  having compact support, the convex hulls of the support of  $u$  and  $f$  are the same. To prove this we only have to show that if  $u$  has its support in some half space  $\xi \cdot x \geq 0$  and  $f$  has its support in a smaller half space  $\xi \cdot x \geq a > 0$ , then indeed  $u$  also has its support there. But this follows at once from (1.2) by setting  $\eta = -\tau\xi$ ,  $\tau > 0$  and letting  $\tau \rightarrow \infty$ , since the right-hand side in (1.2) is  $O(e^{-2\tau a})$ .

This simple device works also in the purely local problem of uniqueness in the Cauchy problem for solutions of

$$(1.3) \quad |Pv| \leq k \sum_{|\alpha| \geq 1} |P^\alpha v|$$

if  $v$  has zero Cauchy data on a hypersurface which is *strictly convex*. If  $v$  is a solution on the convex side  $\mathfrak{D}$ , it vanishes identically near the surface. (Observe that the hypersurface may contain characteristic points.) To prove this, suppose, for simplicity, that the surface is that of a sphere, and that the origin lies on the surface.

We wish to prove that  $v \equiv 0$  in some neighborhood of the origin (having extended  $v$  as zero on the nonconvex side of the surface). We may suppose that the inner-directed unit normal to the surface at the origin is  $(1, 0, \dots, 0)$ . Let  $\zeta(x^1) \geq 0$  be a  $C^\infty$  function for  $x^1 \geq 0$  which equals 1 for  $x^1 < \epsilon$  and vanishes for  $x^1 > 2\epsilon$ ,  $\epsilon$  being small. For  $\epsilon$  small because of the convexity, the set of points in  $\mathfrak{D}$  with  $x_1 \leq 2\epsilon$  may be contained in a sphere of arbitrarily small radius  $r$ . If we now apply (1.2) to the function  $u = \zeta v$  with  $\eta = -\tau(1, \dots, 0)$  we find, with  $c_1(\epsilon)$  independent of  $\tau > 0$ ,

$$\sum_{|\alpha| \geq 1} \|e^{-\tau x^1} \zeta P^{(\alpha)} v\|^2 \leq c r^2 \|e^{-\tau x^1} \zeta P v\|^2 + c_1(\epsilon) e^{-\tau \epsilon}$$

The last term arises from the terms involving derivatives of  $\zeta$ , but these are non-zero only for  $x^1 > \epsilon$ . If we now insert (1.3) in the right-hand side we find, with some constant  $c_2$  independent of  $\tau$ ,

$$\sum_{|\alpha| \geq 1} \|e^{-\tau x^1} \zeta P^{(\alpha)} v\|^2 \leq c_2 r^2 \sum_{|\alpha| \geq 1} \|e^{-\tau x^2} \zeta P^{(\alpha)} v\|^2 + c_1(\epsilon) e^{-\tau \epsilon}$$

Thus if  $\epsilon$  is sufficiently small so that  $c_2 r^2 < \frac{1}{2}$  we have

$$\sum_{|\alpha| \geq 1} \|e^{-\tau x^1} \zeta P^{(\alpha)} v\|^2 \leq 2c_1(\epsilon) e^{-\tau \epsilon}$$

In particular then, restricting integration on the left to  $x^1 < \epsilon/2$  we see that this is possible only if  $v \equiv 0$  for  $x^1 < \epsilon/2$ .

The idea of using inequalities with weight functions, as (1.2), in proving uniqueness for the Cauchy problem goes back to Carleman.

John [26] has observed that (1.2) may be used also to prove regularity of a solution  $v$  of

$$Pv = f, \quad f \in C^\infty$$

in such a domain bounded by a strictly convex surface which is non-characteristic and on which  $v$  has vanishing Cauchy data; Malgrange [38] also proved this. Hörmander [25] has given the most elegant proof of this using (1.2), which proceeds as follows. Using the preceding description, with  $(1, 0, \dots, 0)$  normal to the surface at the origin, we will prove that  $v \in C^\infty$  in some fixed neighborhood  $x^1 < \epsilon$  of the origin assuming, for simplicity, that  $v \in C^m$ . Set  $u = \zeta v$  as before,  $Lu = g$ , and observe that  $g \in C^\infty$  for  $x^1 < \epsilon$  and  $g \in C^1$  for  $x^1 \geq \epsilon$ .

Denote by  $\bar{u}(x^1, \sigma)$ ,  $\sigma = (\sigma_2, \dots, \sigma_n)$  the Fourier transform of  $u$  in the variables  $(x^2, \dots, x^n)$ ,

$$\bar{u}(x^1, \sigma) = \int u(x) \exp \left( -i \sum_2^n x^j \sigma_j \right) dx^2 \cdots dx^n$$

then

$$\bar{g} = P(D_1, \dots, D_n)u = P(D_1, \sigma_2, \dots, \sigma_n)\bar{u}$$

Since the boundary is noncharacteristic at the origin  $(\partial/\partial \xi_1)^m P(\xi) \neq 0$ ; set  $(\partial/\partial \xi_1)^j P(\xi) = P^j$ . We now apply (1.2) for  $\sigma$  fixed to  $\bar{u}$  as a function of the single variable  $x^1$ . Then

$$\int e^{-2\tau x^1} (|\bar{u}|^2 + \sum |P^j \bar{u}|^2) dx^1 \leq c' r^2 \int |e^{-\tau x^1} \bar{g}|^2 dx^1$$

for arbitrary real  $\tau$ . We may choose  $\tau$  as an arbitrary function of  $\sigma$ ; with  $\sigma^2 = \sum \sigma_j^2$  set  $e^{2\tau} = (1 + \sigma^2)^p$ , for some  $p > 0$ . On multiplying both sides of the resulting inequality by  $(1 + \sigma^2)^{p\epsilon}$  we find

$$(1.4) \quad \int (1 + \sigma^2)^{p(\epsilon - x^1)} (|\bar{u}|^2 + \sum_j |P^j \bar{u}|^2) dx^1 \\ \leq c' r^2 \int (1 + \sigma^2)^{p(\epsilon - x^1)} |\bar{g}|^2 dx^1$$

Now observe that since  $g$  belongs to  $C^\infty$  for  $x^1 < \epsilon$  and to  $C^m$  for  $x^1 \geq \epsilon$ , the integral of the right-hand side with respect to  $\sigma$  is convergent (by Parseval) for any  $p > 0$ . Hence this is true also for the term on the left, that is,

$$\iint (1 + \sigma^2)^{p(\epsilon - x^1)} (|\bar{u}|^2 + \sum |P^j \bar{u}|^2) dx^1 d\sigma < \infty$$

Thus in particular, restricting integration to  $x^1 < \epsilon/2$ , where  $u = v$ , we have

$$\iint_{x^1 < \epsilon/2} (1 + \sigma^2)^{p\epsilon/2} (|\bar{v}|^2 + \sum |P^j \bar{v}|^2) dx^1 d\sigma < \infty$$

But since  $p > 0$  is arbitrary, this means (by Parseval) that in  $x^1 < \epsilon/2$  the function  $v$  and its  $x^1$  derivatives up to order  $m - 1$  have square integrable derivatives in the variables  $(x^2, \dots, x^n)$  of all orders. The differential equation  $Pv = f$  enables us to express  $(\partial/\partial x^1)^m v$  in terms of these other derivatives of  $f$ , and hence it too has square integrable derivatives in the variables  $(x^2, \dots, x^m)$  of all orders. By repeated differentiations of the equation we may express further  $x^1$  derivatives in terms of derivatives which we know are square integrable, and thus we find that  $v$  has square integrable

derivatives in all variables of all orders for  $x^1 < \epsilon$ . But then  $v \in C^\infty$ .

Observe that the procedure is to prove first the regularity of the solution in directions tangential to the boundary and then to infer regularity in all variables. This approach is used repeatedly in proving regularity for various boundary value problems.

We should remark that Treves [52] (see also [53]) has given an extension of Hörmander's inequality which is also more useful: for functions  $u$  of compact support,

$$(1.5) \quad t^{2\alpha} \left\| \exp \left( \sum_j (t_j x^j)^2 \right) P^{(\alpha)} u \right\|^2 \leq \text{constant} \left\| \exp \left( \sum_j (t_j x^j)^2 \right) P u \right\|^2$$

where the constant depends only on  $m$  and is independent of the support of  $u$ .

## 2. EQUATIONS WITH VARIABLE COEFFICIENTS

We begin with the question of existence of solutions of equations

$$(2.1) \quad P(x, D)u = f$$

with (complex) variable coefficients. It is of course desirable to make the solution unique by finding appropriate well-posed boundary or initial conditions. It is also of interest, however, to know whether any solutions exist at all. As we have remarked in the previous section, this is always so in the constant coefficient case with  $f \in C_0^\infty$ . Indeed for constant coefficients the equation has been studied in great detail for many classes of functions  $f$ .

As we have remarked, the existence of solutions is intimately tied up with estimates. For instance suppose that  $P^*$  is the formal adjoint of the operator  $P$ , defined in terms of the  $L_2$  scalar product  $(\cdot, \cdot)$  by

$$(Pu, v) = (u, P^*v)$$

where one of the functions  $u, v$  has compact support. If we can establish an a priori estimate of, say, the form—for all functions  $u \in C_0^\infty(\mathfrak{D})$  (or satisfying some less restrictive boundary conditions)

$$(2.2) \quad \|u\|_s \leq C \|P^*u\|_k$$

for some  $s$  and  $k$  and constant  $C$ —then this implies by a functional analytic argument using the Hahn-Banach theorem that  $Pu = f$  always has distribution solution  $u$  in  $\mathfrak{D}$  (satisfying in some sense

adjoint boundary conditions) for every  $f \in C_0^\infty(\mathfrak{D})$ —actually for a wider class of functions  $f$ . Conversely one sees [25] with the aid of the closed graph theorem that existence of distribution solutions of (2.1) in  $\mathfrak{D}$  for every  $f \in C_0^\infty(\mathfrak{D})$  implies the following inequality: if  $\Omega$  is an open subset of  $\mathfrak{D}$  with compact closure in  $\mathfrak{D}$ , there are constants  $C, k, N$  such that

$$(2.3) \quad |(f, v)| \leq C|f|_k |P^*v|_N \quad \text{for all } f, v \in C_0^\infty(\Omega)$$

There are several techniques for obtaining estimates of the form (2.2). (i) Energy integral methods in which we consider the  $L_2$  scalar product of  $(Pu, Mu)$ , where  $M$  is a differential operator chosen in such a way that after suitable use of Green's theorem (making use of the boundary conditions) we obtain the integral of a positive quadratic  $Q(u, u)$  of  $u$  and some derivatives. Thus

$$|(Pu, Mu)| \geq Q(u, u) + B(u, u)$$

where  $B(u, u)$  is a similar boundary integral which is definite in virtue of the boundary conditions. (ii) Fourier transform and Parseval's theorem; this may also be used in proving positive definiteness of integrals of quadratic forms arising in (i). (iii) Integral representation of solutions via integral operators—often obtained from the fundamental solution or an approximate fundamental solution.

It was believed that every differential equation (2.1) with  $f \in C^\infty$  possessed solutions, at least locally, but in 1957 Lewy [35] presented a first-order operator  $P$  (with analytic coefficients) such that  $Pu = f$  has no distribution solution in any open set for most  $C^\infty$  functions  $f$ . Hörmander [23] (see also [25]) then derived necessary conditions for solvability of an operator  $P$ . We say that  $P$  is a *solvable* at a point  $x_0$  if there is an open neighborhood  $\Omega$  of  $x_0$  such that for every  $f \in C_0^\infty(\Omega)$  the equation  $Pu = f$  has a distribution solution in  $\Omega$ . His condition is expressed in terms of the commutator  $[P', \bar{P}']$  of the principal part  $P'$  of  $P$  and its complex conjugate  $\bar{P}'$  obtained by conjugating all coefficients; this commutator is a differential operator of order  $\leq 2m - 1$ . Denote by  $C(x, \xi)$  the characteristic polynomial of its leading part, that is, of terms of the order  $2m - 1$ . Then Hörmander's necessary condition for solvability at  $x_0$  is:

(H) For  $x$  in a neighborhood of  $x_0$ , every real root  $\xi$  of  $P'(x, \xi)$  is also a root of  $C(x, \xi)$ .

In proving the necessity of (H), Hörmander first showed that (2.3) is a necessary condition and then, assuming that (H) does not hold at  $x_0$  he constructed a family of functions (in any given neighborhood  $\Omega$  of  $x_0$ ) depending on a parameter  $\tau$ ,  $f_\tau, v_\tau \in C_0^\infty(\Omega)$  such that as  $\tau \rightarrow \infty$  these functions violated (2.3). The construction of  $v_\tau$ , an approximate solution of  $L^*v = 0$ , is a bit involved.

For first-order operators  $P$ , Nirenberg and Treves [46] have found both further necessary conditions and also sufficient conditions which agree if the leading coefficients of  $P$  are analytic. These conditions may be expressed in terms of the successive commutators  $C_k = [P', C_{k-1}]$ ,  $k = 2, 3, \dots$ ,  $C_1 = [P', \bar{P}']$  which are all first-order operators. At any point  $x$  let  $k_0(x)$  be the first value of  $k$ , if finite, such that  $C_k(x, \xi)$  does not vanish at every real root  $\xi$  of  $P'(x, \xi)$ . Then the necessary and sufficient conditions for  $P$  to be solvable at  $x_0$  is:

(P') At every point  $x$  in some neighborhood of  $x_0$ ,  $k_0(x)$ , if finite, is even.

This condition may also be expressed in terms of the bicharacteristics of the real and imaginary parts of  $P'(x, \xi)$ . Let  $A(x, \xi)$  and  $B(x, \xi)$  be the real and imaginary parts of  $P'(x, \xi)$ .

*Definition.* A bicharacteristic of  $A(x, \xi)$  is a curve in the real  $(x, \xi)$  space satisfying the differential equation system (where  $\cdot$  denotes differentiation along the bicharacteristic)

$$(2.4) \quad \dot{x}^j = \frac{\partial}{\partial \xi_j} A(x, \xi), \quad -\dot{\xi}_j = \frac{\partial}{\partial x^j} A(x, \xi)$$

We see immediately that  $A$  is constant along any bicharacteristic.

Condition (P') (for analytic coefficients) is equivalent to the condition:  $B$  does not change sign on any bicharacteristic of  $A$  on which  $A$  vanishes, and also  $A$  does not change sign on any bicharacteristic of  $B$  on which  $B$  vanishes.

The proof of necessity of (P') is simply an extension of Hörmander's necessity proof, whereas sufficiency is proved by establishing an estimate (2.2) with  $s = 0$ ,  $k = 1$ ; the argument is slightly tricky.

We should remark that an operator may be solvable at every point in a domain without being globally solvable in the domain. An analogous situation maintains even for operators with constant coefficients. Malgrange [37] has shown that the equation  $P(D)u = f$

is always solvable in a domain  $\mathfrak{D}$  for every  $C^\infty$  function  $f$  in  $\mathfrak{D}$  (with  $u$  a distribution in  $\mathfrak{D}$ ) if and only if  $\mathfrak{D}$  is  $P$ -convex. To be  $P$ -convex means that to any compact set  $K_1 \subset \mathfrak{D}$  there is another compact set  $K_2 \subset \mathfrak{D}$  such that any distribution  $u$  with compact support in  $\mathfrak{D}$  and with support  $P(-D)u$  in  $K_1$  has its support in  $K_2$ . (Since the convex hull of  $\text{supp } P(-D)u$  agrees with that of  $u$  it follows that any bounded convex set is  $P$ -convex for every operator  $P$  with constant coefficients.) Another way of saying that  $\mathfrak{D}$  is  $P$ -convex is to say that uniqueness for the Cauchy problem for  $P(-D)$  holds in some neighborhood of the boundary, that is, if  $u$  is defined on an open neighborhood  $\Omega_1$  of the boundary of  $\mathfrak{D}$ , satisfies  $P(-D)u = 0$ , and vanishes outside  $\mathfrak{D}$ , then  $u$  vanishes on a neighborhood  $\Omega_2$  of the boundary, where  $\Omega_2$  depends only on  $\Omega_1$  and not on the particular solution  $u$ .

Hörmander [24] has completely characterized the  $P$ -convex domains for  $n = 2$ , and has given a sufficient condition for  $n > 2$  which we may describe. Assume that  $\partial\mathfrak{D}$  is a  $C^2$  hypersurface. Suppose that at every point where  $\partial\mathfrak{D}$  is characteristic with respect to  $P$  the curvature of the hypersurface in the direction of either  $\text{Re grad } P'(N)$  or  $\text{Im grad } P'(N)$  is positive (here  $N$  is the unit normal to  $\partial\mathfrak{D}$  at the point). Then  $\mathfrak{D}$  is  $P$ -convex.

Thus instead of assuming convexity of the boundary, as in Section 1, we assume only convexity of the boundary along a particular *bicharacteristic* of either  $\text{Re } P'$  or  $\text{Im } P'$ .

The idea of the proof is easily sketched. We must prove uniqueness of the Cauchy problem for  $P(-D)u = 0$  in some neighborhood of the boundary. The argument is local. At a point where the boundary is not characteristic we obtain a neighborhood of the point where  $u \equiv 0$  by Holmgren's theorem. At a point  $x_0$  where  $\mathfrak{D}$  is characteristic, Hörmander makes clever use of the hypothesis to construct a family of hypersurfaces (or rather pieces of hypersurfaces) whose boundaries lie outside  $\mathfrak{D}$  and which together cover a neighborhood of  $x_0$  and such that *none of them contains any characteristic points*. He then employs Holmgren again to conclude that  $u \equiv 0$  in a neighborhood of  $x_0$ . This material is also contained in [25] and [53].

In Chapter 8 of his book [25] Hörmander has obtained much more general results than those derived in Section 1 from (1.2), and has also extended them to operators with variable coefficients. He treats global existence problems and uniqueness and regularity

in the Cauchy problem. In this work the bicharacteristics play a very essential role.

The operators he considers are either elliptic or belong to a class termed principally normal.  $P$  is principally normal if there is a differential operator  $Q(x, D)$  homogeneous of order  $m - 1$  in  $D$  such that the leading part  $C$  of the commutator  $[P', \bar{P}']$  may be expressed as

$$(2.5) \quad C(x, \xi) = P'(x, \xi)\bar{Q}(x, \xi) + Q(x, \xi)\bar{P}'(x, \xi)$$

In view of the necessary condition (H) this is not an unreasonable hypothesis.

In addition, Hörmander introduces the important notions of pseudoconvexity of a hypersurface and strong pseudoconvexity. To explain these notions suppose first that  $P'(x, \xi)$  has real coefficients and suppose that the  $x$  projection of a bicharacteristic curve given by (2.4) on which  $P'$  vanishes is tangent to it at a point  $x_0$ , that is, if the surface is given by  $\psi(x) = \psi(x_0)$  with  $\text{grad } \psi(x_0) \neq 0$ , we suppose

$$(2.6) \quad P'(x_0, \xi) = 0$$

$$\sum_j \frac{\partial}{\partial \xi_j} P'(x_0, \xi) \frac{\partial \psi}{\partial x^j}(x_0) = 0$$

If  $\cdot$  denotes differentiation with respect to the parameter on the bicharacteristic, we see that on the curve  $\psi(x_0) = 0$  whereas (using summation convention)

$$\begin{aligned} \ddot{\psi}(x_0) &= \psi_{x^i x^k} \dot{x}^j \dot{x}^k + \psi_{x^i} \ddot{x}^j \\ &= \psi_{x^i x^k} P'_{\xi_i} P'_{\xi_k} + \psi_{x^i} (P'_{\xi_i x^k} P'_{\xi_k} - P'_{\xi_i \xi_k} P_{x^k}) \end{aligned}$$

in virtue of (2.4). Pseudoconvexity is the requirement that  $\psi$  be a convex function on such a curve at such a point; in fact it requires that  $\ddot{\psi}(x_0) > 0$ . Thus a bicharacteristic tangent to the surface must lie on the side where  $\psi > \psi(x_0)$ . More generally for complex coefficients the definition is as follows:

*Pseudoconvexity.* An oriented hypersurface  $\psi(x) = \psi(x_0)$  is pseudoconvex with respect to  $P$  at a point  $x_0$  if at  $x_0$

$$(2.7) \quad \psi_{x^i x^k} P'_{\xi_i} \bar{P}'_{\xi_k} + \psi_{x^i} \text{Re} (P'_{\xi_i x^k} \bar{P}'_{\xi_k} - \bar{P}'_{\xi_i \xi_k} P_{x^k}) > 0$$

for every real  $\xi \neq 0$  satisfying (2.5).

He then proves the following global existence theorem (we are omitting the precise smoothness condition on the functions).

*Theorem 1.* Suppose that  $P$  is a principally normal operator in a domain  $\mathfrak{D}$  in which there is a real  $C^2$  function  $\psi$  with  $\text{grad } \psi \neq 0$ , all of whose level surfaces are pseudoconvex. Let  $\Omega$  be an open set with compact closure in  $\mathfrak{D}$  and assume that  $f$  is a function in  $\mathfrak{D}$  which is orthogonal in the  $L_2$  sense to all solutions  $v$  of the adjoint problem  $P^*v = 0$  having compact support in  $\Omega$ . Then there is a solution in  $\Omega$  of  $Pu = f$ . Furthermore if  $f \in C^\infty$  there is a  $C^\infty$  solution.

In addition he proves the following regularity theorem in the Cauchy problem.

*Theorem 2.* Suppose that  $P$  is principally normal and that  $\psi$  is defined in a neighborhood of  $x_0$ ,  $\text{grad } \psi \neq 0$ , with its level surface through  $x_0$  pseudoconvex there. If  $u$  is a distribution solution of  $Pu = f$  in a neighborhood of  $x_0$  with  $f \in C^\infty$ , and  $u \in C^\infty$  on the side of the surface where  $\psi > \psi(x_0)$ , then  $u \in C^\infty$  in a whole neighborhood of  $x_0$ .

The first theorem is proved by deriving estimates similar to the estimates (1.2), but with weight factor  $e^{\phi}$  where  $\phi$  is a nonlinear function closely related to  $\psi$ . The second theorem is proved by a suitable analogue of (1.5). The details of the proof are too long to be described here.

We remark that the pseudoconvexity condition in the second theorem is essential. Following a construction of Zerner [55], Hörmander [25] constructs a solution of an equation  $Pu = 0$  with constant coefficients which is  $C^\infty$  everywhere except on a bicharacteristic.

Strong pseudoconvexity imposes an additional condition at the nonzero double roots in  $\tau$  of  $P(x_0, \xi + i\tau \text{ grad } \psi(x_0))$ . Assuming strong pseudoconvexity he proves uniqueness for the Cauchy problem near  $x_0$  for zero data on the surface  $\psi = \psi(x_0)$ . This result generalizes the previous results by Calderon [8] and by Hörmander [22] in which double roots are excluded. Examples of Cohen [11] show that the condition cannot be omitted. The method of proof yields not only uniqueness in the Cauchy problem but also estimates for the solution. In general, of course, a uniqueness theorem should always provide estimates for the solution. We

remark that John [26] has adapted the Holmgren uniqueness theorem for analytic equations so as to yield estimates for the solution.

### 3. BOUNDARY VALUE PROBLEMS

Up to now we have considered differential equations without imposing boundary conditions. For equations of well-determined "type," elliptic, hyperbolic, etc. boundary value problems have been studied over many years. Equations of arbitrary (or unknown) "type," however, have only been considered recently. Friedrichs [16] has developed an existence theory of boundary value problems for first-order systems which are symmetric and positive, that is, which have the form (real coefficients):

$$(3.1) \quad Lu = \left( \sum A^j \frac{\partial}{\partial x^j} + A \right) u = f$$

where  $u = (u^1, \dots, u^N)$  is a system of  $N$  real functions,  $f = (f^1, \dots, f^N)$ , the  $A^j$  and  $A$  are square matrices, the  $A^j$  being symmetric; the positivity is expressed by the requirement that the matrix

$$(3.2) \quad K = A - \frac{1}{2} \sum \frac{\partial}{\partial x^j} A^j$$

have positive definite symmetric part. Friedrichs' results have been extended and simplified by Lax and Phillips [34] (see also references there to Phillips), and further results have been obtained by Sarason [48].

Before describing the boundary conditions we observe that, by the symmetry of the  $A^j$  (using summation convention),

$$uLu = \frac{1}{2}(uA^j u)_{x^j} + uKu$$

Thus by Green's theorem

$$(3.3) \quad (Lu, u) = (Ku, u) + \int_{\mathfrak{D}} u A_n u \, ds$$

where  $\mathfrak{D}$  denotes the boundary of the domain  $\mathfrak{D}$  and  $ds$  denotes its element of area;  $A_n = n_j A^j$ , where  $n_j$  denotes the  $j$ th component of the outer normal at the boundary.

In describing the boundary conditions we suppose that the boundary  $\mathfrak{D}$  is decomposed into several smooth parts  $\Gamma_1, \dots, \Gamma_k$ , and that on each  $\Gamma_j$  there is a linear subspace  $N_j(x)$  varying smoothly with  $x$  of fixed dimension  $m_j$ . We then require that for  $x$  in  $\Gamma_j$  the boundary values of  $u$  lie in  $N_j(x)$ . These boundary conditions are called positive if at every point  $x$  of  $\Gamma_j$  the operator  $A_n$  is positive over  $N_j(x)$ , that is,  $uA_nu \geq 0$  for all  $u(x)$  in  $N_j(x)$ ,  $j = 1, \dots, k$ . We see from (3.3) that if  $u$  satisfies positive boundary conditions and if the system is symmetric positive, the following energy estimate holds.

$$(3.4) \quad (u, u) \leq \text{constant} (Lu, u)$$

To solve such a boundary value problem we should ensure that there are not too many boundary conditions (for we may be able to enlarge some space  $N_j$  and still maintain positivity). Thus we assume that each  $N_j$  is maximal positive (so in particular it contains the null space of  $A_n$  at each point  $x \in \Gamma_j$ ). It then follows that the adjoint boundary conditions are positive for the adjoint operator  $L^*$ . These boundary conditions are expressed by the requirement that  $u$  lie in  $P_j(x)$  for  $x \in \Gamma_j$  where  $P_j$  is the orthogonal complement of  $A_n(x)N(x)$ ; clearly we have

$$(Lu, v) = (u, L^*v)$$

if  $u$  and  $v$  satisfy the boundary and adjoint boundary conditions. Note that the dimensions  $m_j$  depend strongly on the nature of the operator  $A_n$ .

By using the projection theorem in Hilbert space, it is not difficult to establish the existence of weak solutions of such symmetric positive problem with positive boundary conditions. But weak solutions are not necessarily unique, and the real work comes in establishing the regularity of the solutions. Uniqueness for regular solutions follows from (3.4). Under an additional condition Friedrichs proved sufficient regularity of the weak solutions to guarantee uniqueness. Lax and Phillips [34] have presented a simpler and more general regularity theorem near the boundary for rather general first-order systems not necessarily symmetric (see also [48]). The argument is local; by first smoothing the solution in directions tangent to the boundary they demonstrate regularity in the tangential directions and then derive general regularity.

Friedrichs has shown that many problems of mathematical physics may be expressed by symmetric positive systems (although often it is not trivial to do this) and may be treated by his theory. For instance he treats the Tricomi problem involving an equation which is elliptic in some region and hyperbolic in another.

This situation in which the main difficulty is proving regularity of the solution is typical. The general problem of finding under which boundary conditions solutions of a given differential equation are necessarily regular has not been completely solved. The case of a single equation with constant coefficients for one function in a half space restricted on the boundary by differential boundary conditions, also having constant coefficients, has been solved by Hörmander [21]. In general the results will depend on the lower-order terms (not just the highest) and on the nature of the boundary (see the problem of Section 7).

#### 4. SINGULAR INTEGRAL OPERATORS

We conclude this lecture by describing the construction and basic properties of a class of operators introduced by Calderon, Zygmund [10], and Mihlin [39]. These include differential operators, in the sense that any differential operator may be expressed in terms of these and the Laplace operator. These operators have already proven to be a very useful and important tool, and will continue to be very useful. For further references we may mention also [9], [51], [50], and [33]. We shall restrict ourselves to operators on scalar valued square integrable functions defined in all of space.

The operators are a special algebra of operators mapping the space of  $L_2$  functions continuously into itself (we may also consider  $H_s$  into itself), which are considered only modulo the ideal  $K$  of compact operators. Furthermore they are local operators in the sense that if  $T$  is such an operator and if  $u, v$ , are  $L_2$  functions with supports in fixed disjoint sets then  $(u, Tv)$  is compact in  $u$  and in  $v$ .

We start with the following special operators.

1. Multiplication by a  $C^\infty$  function  $a(x)$  having a limit at infinity.
2. Multiplication of the Fourier transform  $\hat{u}(\xi)$  of  $u(x)$ ,

$$\hat{u}(\xi) = (2\pi)^{-n/2} \int e^{ix \cdot \xi} u(x) dx = Fu$$

by a function  $\alpha(\xi)$  which is homogeneous of degree zero and  $C^\infty$  on the unit sphere  $|\xi| = 1$ .

These operators are clearly continuous maps of  $L_2$  into itself. We shall also have need of the operator  $\wedge$ , the square root of the Laplace operator, which may be defined, in terms of the Fourier transform operator  $F$  by

$$(4.1) \quad \wedge = F^{-1}|\xi|F$$

Here  $|\xi|$  denotes multiplication by  $|\xi|$ . The operator  $\wedge$  is defined on a dense set in  $L_2$ , for instance the  $C^\infty$  functions dying down rapidly at infinity. Let us in fact confine ourselves to these for the moment.

The operators of 1 and 2 which we denote simply by  $a, \alpha$  have the important property that their commutators are compact operators. In fact the operators

$$(4.2) \quad \wedge[a, \alpha], \quad [a, \alpha]\wedge$$

are bounded as maps of  $L_2$  into  $L_2$ .

The algebra of operators we will consider will be generated by the special operators of the type 1 and 2. Consider first finite sums

$$H = \sum_1^N a_j \cdot \alpha_j$$

With such a sum we associate the *symbol*

$$h(x, \xi) = \sum a_j(x)\alpha_j(\xi)$$

According to the preceding these form an algebra modulo  $K$ . Naturally we want to complete this algebra by taking its closure in the norm sense. We would expect that to every "symbol," that is, a function  $h(x, \xi)$  which is homogeneous of degree zero in  $\xi$  and sufficiently regular say of class  $C^k$ ,  $k = k(n)$  in  $(x, \xi)$  (for  $|\xi| = 1$ ), and having a limit (depending on  $\xi$ ) as  $x \rightarrow \infty$ , it is possible to define a corresponding bounded operator  $H$  of  $L_2$  into  $L_2$  mod  $K$ . Indeed this is the case for  $k$  sufficiently large and may be shown by an expansion of  $h$  in the form  $h = \sum a_j(x)\alpha_j(\xi)$ . Furthermore, the commutator of any two operators  $G, H$  obtained in this way is again compact, and in fact  $\wedge[G, H], [G, H]\wedge$  are bounded operators.

We may now represent any differential operator (with smooth coefficients having limiting values at  $\infty$ ) in a special way. Suppose the operator is given in the form  $P = \sum a_\alpha D^\alpha$ . The operator  $D^\alpha$

may be written as

$$D^\alpha = F^{-1} \xi^\alpha F = F^{-1} \left( \frac{\xi}{|\xi|} \right)^\alpha F \wedge^{|\alpha|} = \Xi^{(\alpha)} \wedge^{|\alpha|}$$

where  $\Xi^{(\alpha)}$  is here defined. Then we may write

$$\begin{aligned} P &= \sum a_\alpha \Xi^{(\alpha)} \wedge^{|\alpha|} \\ &= \sum \wedge^{|\alpha|} a_\alpha \Xi^{(\alpha)} \pmod{K} \end{aligned}$$

in virtue of the preceding. If  $P'$  is the principal part of  $P$ , of order  $m$ , then  $P' = \sum a_\alpha \Xi^{(\alpha)} \wedge^m$ ; thus we associate to  $P$  the singular integral operator

$$\sum a_\alpha \Xi^{(\alpha)} \text{ with symbol } \sum a_\alpha(x) \xi^\alpha = P'(x, \xi)$$

Returning to the algebra of operators obtained from (smooth) symbols,  $h(x, \xi)$  it is shown that the mapping: algebra of symbols  $\rightarrow$  algebras of operators (mod  $K$ ), is one-to-one and so an algebraic homomorphism. Also for any smooth function  $u(x)$  dying down rapidly at  $\infty$  we have the representation

$$(Hu)(x) = (2\pi)^{-n} \int e^{ix\xi} h(x, \xi) \bar{u}(\xi) d\xi$$

and the representation as a singular integral operator with a kernel  $k$ :

$$(Hu)(x) = \int k(x, x-y) u(y) dy$$

where  $k(x, z)$  is homogeneous in  $z$  of degree  $-n$  and regular in  $(x, z)$ , for  $|z| = 1$ . Furthermore, the norm of the operator  $H$  (mod  $K$ ) turns out to be  $\sup |h(x, \xi)|$ . Therefore, taking closure we see that we can assign to any *continuous* symbol  $h(x, \xi)$  an operator (mod  $K$ ) (i.e., an equivalence class of bounded operators of  $L_2 \rightarrow L_2$  modulo  $K$ ).

*Elliptic Operator.* We observe that if a symbol  $h(x, \xi)$  does not vanish, the corresponding operator  $H$  has an inverse  $H^{-1}$  (mod  $K$ ) defined by the symbol  $h^{-1}(x, \xi)$ . Thus such an operator  $H$  has a finite dimensional null space. Furthermore, its range is closed and has finite codimension. (The converse is also true; these properties of the null space and the range imply that  $h$  does not vanish.) Hence it has a finite index:

$$\text{Index} = \text{dimension of null space} - \text{codimension of range}$$

In analogy with differential operators the corresponding operator is called elliptic.

The representation of differential operators by means of singular integral operators was used by Calderon in his study of the Cauchy problem [8] and in treating hyperbolic equations [9]. They have also been used recently in elliptic problems (see Section 8) where suitable extensions to manifolds are used. More recently, in a very interesting paper,\* Calderon has proved, with their aid, existence theorems in the large for very general systems of partial differential equations. The demonstrations involve new uniqueness theorems for the Cauchy problem in a half-space.

## *Lecture II Elliptic Differential Equations in Differential Geometry*

We shall take up three geometric-analytic topics that are concerned with or make use of recent developments in elliptic partial differential equations. These are:

§6. Work of Kodaira, Spencer, and Kuranishi on existence and deformation of complex structure on compact manifolds.

§7. The  $\bar{\partial}$ -Neumann problem; work of Kohn and Morrey.

§8. The index problem for elliptic operators on compact manifolds; work of Atiyah, Singer.

In Section 6 we describe the problems and attempt to give the structure of the arguments used. In the other sections the results and techniques are too involved and we can only give the merest hint about the nature of the analytic methods employed.

### 5. PRELIMINARIES

As a well-known illustration of the use of elliptic equations in global differential geometry, let us recall some basic facts of the Hodge-Kodaira theory. Let  $V$  be a compact Riemannian manifold. The Riemannian metric enables one to define  $\delta$  the adjoint of the exterior differential operator  $d$  acting on exterior differential forms.

\* A. P. Calderon, "Existence and uniqueness theorems for systems of partial differential equations," Fluid Dynamics and Applied Mathematics Proceeding of Symposium at Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, April 1961. Gordon and Breach (1962) 147-195.

The operator  $d + \delta$  on all forms is elliptic. Also the Laplace operator

$$\Delta = d\delta + \delta d$$

which preserves the degree is strongly elliptic, and one has the Hodge-Kodaira decomposition theorem: there exists a Green's operator  $G$  commuting with  $d$  and  $\delta$  such that

$$I = H + \Delta G = H + d\delta G + \delta dG$$

where  $H$  is the projection onto the finite dimensional space of harmonic forms  $\phi$  (solutions of  $\Delta\phi = 0$ ). Thus any form can be decomposed uniquely into a harmonic form  $+d$  of a form  $+\delta$  of a form. (Furthermore the space of harmonic forms of degree  $p$  has finite dimension equal to the  $p$ th Betti number of  $V$ .) This decomposition formula, derived by Hilbert space arguments, may be extended in a similar way to any strongly elliptic operator or system of the form

$$\tilde{\Delta} = \tilde{d}\tilde{\delta} + \tilde{\delta}\tilde{d}$$

if  $\tilde{d}^2 = 0$  and  $\tilde{\delta}$  is the adjoint of  $\tilde{d}$ .

We turn now to some technical material on elliptic boundary value problems which is used in the subsequent sections. Consider an elliptic system of  $N$  equations for  $N$  functions  $u = (u_1, \dots, u_N)$  of  $n$  variables  $x = (x^1, \dots, x^n)$ . We use summation convention

$$(5.1) \quad l_{ij}(D)u_j = f_i, \quad i = 1, \dots, N$$

where the  $l_{ij}$  are differential operators of, for simplicity (see however [13] and [2], part II), the same order  $m$ . Ellipticity means that the characteristic matrix  $l'_{ij}(\xi)$  of leading terms is nonsingular for real  $\xi = (\xi_1, \dots, \xi_n) \neq 0$ ; if  $n > 2$ , this implies that  $mN$  is even, and we shall always assume this to be the case. If the quadratic form associated with the characteristic matrix is positive definite for real  $\xi \neq 0$  the system is called strongly elliptic. For functions  $u$  defined in a bounded domain we also impose  $mN/2$  differential boundary conditions

$$(5.2) \quad B_{\alpha j}(D)u_j = \phi_{\alpha}, \quad \alpha = 1, \dots, \frac{mN}{2}$$

of orders  $< m$  (for simplicity).

The boundary conditions are said to be coercive in some norm  $\| \cdot \|$  [such as  $L_2$ ,  $L_p$ , or  $\| \cdot \|_s$  norm, or Hölder norm  $\| \cdot \|_\alpha$  defined by  $\|f\|_\alpha = \sup |f| + \sup (|f(x) - f(y)|/|x - y|^\alpha)$ ] if the following norms are equivalent (where  $\| \cdot \|_\alpha$  are norms for functions defined on the boundary):

$$(5.3) \quad \sum_j \sum_{|\alpha| \leq m} \|D^\alpha u_j\|; \quad \sum_i \|l_{ij} u_j\| + \sum_j \|u_j\| + \sum_\alpha \|B_{\alpha j} u_j\|_\alpha$$

Then we can estimate all the derivatives occurring in  $l_{ij} u_j$  in terms of the  $l_{ij} u_j$ . It is easy to estimate all such derivatives in compact subsets of the domain, but these estimates hold up to the boundary if and only if the boundary operators satisfy known algebraic conditions. This problem has been treated in detail for various norms by Agmon, Douglis, and Nirenberg [2], and Browder [7]. The more general question, initiated by Aronszajn, when such estimates hold in  $L_2$  for any number of differential operators (not just  $N$ ) and any number of boundary conditions, has been solved by Agmon [1] and Hörmander for  $N = 1$  and by Figueiredo [14] and Hörmander for arbitrary  $N$ . Hörmander's book [25] also contains a treatment of the existence theory. Lions and Magenes [36] (where other references may be found) have made extensions of the results of [2] using the theory of interpolation of operators in Banach space; see also Schechter [49]. Many recent developments in elliptic theory are contained in the lecture notes [47] by Peetre, where other references to existence theory can be found.

Under the conditions for coerciveness the elliptic operator (with these boundary conditions) has finite dimensional  $s$  null space, and closed range with finite codimension  $r$ , so that the index  $s - r$  is well defined. Furthermore, if the spectrum is not the whole plane it is discrete. The estimates described here are used in Sections 6 and 8. In particular, in Section 6 where a nonlinear problem occurs, the estimates with special norms play an important role. The spaces with these norms are such that the products of two functions in the space belongs again to the space. This is true for the norms  $\| \cdot \|_{k+\alpha}$ , that is, derivatives up to order  $k$  satisfying a Hölder condition with exponent  $\alpha$ , and for the spaces  $H_s$ , for  $s > n/2$ .

Certain boundary value problems arise in practice where the conditions for coerciveness are not satisfied, but where nevertheless the null space is finite dimensional and the range is closed. In such a case the lower-order terms, or the geometric nature of the bound-

ary, such as pseudoconvexity, plays an essential role. The  $\bar{\partial}$ -Neumann problem of Section 7 is a problem of this kind.

## 6. EXISTENCE AND DEFORMATION OF COMPLEX STRUCTURE ON MANIFOLDS\*

We shall give a brief description of the use of differential equations in some of the work of Kodaira, Spencer, and Kuranishi on deformation of complex structure on a compact manifold. Some of the questions they considered grew out of work by Frölicher and Nijenhuis. Throughout,  $V$  will be used to denote a compact  $C^\infty$  manifold. In addition to the papers cited a useful reference is [17].

We recall that a manifold  $V$  (of even dimension  $2n$ ) is said to be a complex manifold if it is possible to cover  $V$  by coordinate patches in each of which the local  $2n$  real coordinates may be expressed as distinguished complex coordinates  $z^1, \dots, z^n$  in such way that the transformation between local coordinates of overlapping patches is expressed by holomorphic functions; for  $n = 1$  such a manifold is simply a Riemann surface. The space of first-order forms with complex coefficients may then be decomposed into the two spaces: the forms spanned by  $dz^1, \dots, dz^n$ , called type  $(1, 0)$  and that spanned by the complex conjugates  $d\bar{z}^1, \dots, d\bar{z}^n$ , type  $(0, 1)$ , these spaces being invariant under (holomorphic) change of coordinates. Furthermore, the following space of forms, called type  $(p, q)$ , is invariant; forms spanned by exterior products of  $p$  forms of type  $(1, 0)$  and  $q$  forms of type  $(0, 1)$ . The exterior differentiation operator  $d$  then decomposes naturally into  $d = \partial + \bar{\partial}$  where

$$\begin{aligned} \partial(a \, dz^{r_1} \wedge \dots \wedge dz^{r_p} \wedge d\bar{z}^{s_1} \wedge \dots \wedge d\bar{z}^{s_q}) \\ = \sum_k \frac{\partial a}{\partial z^k} dz^k \wedge dz^{r_1} \wedge \dots \wedge d\bar{z}^{s_q} \\ \bar{\partial}(a \, dz^{r_1} \wedge \dots \wedge dz^{r_p} \wedge d\bar{z}^{s_1} \wedge \dots \wedge d\bar{z}^{s_q}) \\ = \sum_k \frac{\partial a}{\partial \bar{z}^k} d\bar{z}^k \wedge dz^{r_1} \wedge \dots \wedge d\bar{z}^{s_q} \end{aligned}$$

$\partial$  and  $\bar{\partial}$  map forms of type  $(p, q)$  into type  $(p+1, q)$  and  $(p, q+1)$  respectively, they satisfy  $\partial^2 = \bar{\partial}^2 = 0$  and  $\partial\bar{\partial} + \bar{\partial}\partial = 0$ .

\* The presentation here is based on several conversations with Prof. M. Kuranishi to whom the author wishes to express his warm thanks.

On a complex manifold we may introduce a real analytic Riemannian metric which is also Hermitian, that is, which in terms of distinguished holomorphic coordinates is given by  $g_{j\bar{k}} dz^j d\bar{z}^k$  (using summation convention). Such a metric may be used to define, in the usual way, an  $L_2$  scalar product  $(\ , \ )$  for forms of the same type, and, correspondingly the formal adjoint  $\vartheta$  of  $\delta$  by  $(\vartheta\phi, \psi) = (\phi, \delta\psi)$ ;  $\vartheta$  sends forms of type  $(p, q)$  into  $(p, q - 1)$  and  $\vartheta^2 = 0$ . We readily verify that the operator

$$(6.1) \quad \square = \delta\vartheta + \vartheta\delta$$

preserves type, commutes with  $\delta$  and  $\vartheta$ , and is a (strongly) elliptic partial differential operator on forms of any fixed type.

A given manifold  $V$  may have different complex analytic structures put on it which are not equivalent. Letting  $V_1, V_2$  denote  $V$  with two analytic structures on it, we say that  $V_1$  and  $V_2$  are equivalent if there is a diffeomorphism of  $V_1$  onto  $V_2$  which is also holomorphic with respect to the analytic structures. Since Riemann, people have studied deformation of the complex structure on a Riemann surface; the (complex) parameters on which the deformations depend are called the moduli, and Riemann found the number of moduli of a compact surface of genus  $p$  to be the dimension of the Lie group of complex analytic automorphisms of the surface plus  $3p - 3$ . In recent years, after Teichmüller's work on extremal quasiconformal mappings, the global study of the space of moduli which, after some normalization, is called the Teichmüller space, has been carried out by various authors—Rauch, Ahlfors, Bers, Weil, and others. We refer the reader to [4] and [6] for a description of this work and references.

The problems treated in Kodaira and Spencer [28], and Kuranishi [32] (and a paper to appear), which we shall discuss, are the following:

1. Given a complex analytic manifold  $V_0$ , can we construct a nontrivial family  $V_t$  of complex structures on  $V_0$  (with  $V$  as basic underlying  $C^\infty$  structure) depending continuously on a finite number of parameters  $t$  in a set  $M$  in a Euclidean  $t$ -space  $T$ , with  $V_0$  corresponding to  $t = 0$ ? Can we take  $M$  to be a subvariety and require the dependence on  $t$  to be  $C^\infty$  (or even analytic)? We say that the family  $V_t$  is centered at  $V_0$ .

Consider another family  $\tilde{V}_s$  centered at  $V_0$  depending on param-

an  $L_2$  scalar product on  $\Phi^{0,q}$ , which we denote again by  $(\ , \ )$ . The operator  $\bar{\partial}$  sends  $\Phi^{0,q}$  into  $\Phi^{0,q+1}$ , and its formal adjoint we denote again by  $\partial$ ; the operator  $\square = \partial\bar{\partial} + \bar{\partial}\partial$  is again seen to be strongly elliptic, and we may therefore obtain the analogue of the Hodge-Kodaira decomposition theorem: There is an operator  $G$ , the Green's operator, commuting with  $\bar{\partial}$  and  $\partial$ , such that the identity operator may be decomposed into

$$(6.7) \quad I = H + \partial\bar{\partial}G + \bar{\partial}\partial G$$

where  $H$  is the orthogonal projection (in terms of our scalar product) into the harmonic space, that is, the space of forms annihilated by  $\square$ . This space has finite dimension, say  $m_q$ , which, it should be noted, is independent of the choice of Hermitian metric. Thus any  $\phi \in \Phi^{0,q}$  admits the unique orthogonal decompositions into

$$(6.8) \quad \phi = H\phi + \partial\phi_1 + \bar{\partial}\phi_2$$

Note that  $\square\phi = 0 \Leftrightarrow \bar{\partial}\phi = 0$  and  $\partial\phi = 0$ , by Green's theorem.

We shall show, following Kuranishi [32] (and an unpublished paper) that the set of solutions  $\omega \in \Phi^{0,1}$  of a larger system

$$(6.9) \quad \begin{aligned} \bar{\partial}\omega &= [\omega, \omega] \\ \partial\omega &= 0 \end{aligned}$$

near  $\omega = 0$  may be described in terms of  $m_1$  complex parameters  $t = (t_1, \dots, t_{m_1})$  lying on a complex analytic set, and that this set yields a locally complete family  $V_t$  of complex structures.

We observe that the linearized problem

$$\bar{\partial}\omega = 0, \quad \partial\omega = 0$$

has an  $m_1$  (complex) dimensional space of solutions, namely the elements of  $H\Phi^{0,1}$ . This suggests that the nonlinear problem should also have a finite parameter family of solutions near  $\omega = 0$  which might be obtainable by a bifurcation process (see, for example, Section 7.2 in [44]).

Any solution of (6.9) satisfies, in virtue of (6.7),

$$\omega = H\omega + \partial G[\omega, \omega]$$

By a slight modification of the standard bifurcation process we shall solve the nonlinear equation

$$(6.10) \quad \omega = \omega_0 + \partial G[\omega, \omega]$$

where  $\omega_0$  is an arbitrarily prescribed element of  $H\Phi^{0,1}$ . If  $\phi_1, \dots, \phi_{m_1}$  is a fixed basis for  $H\Phi^{0,1}$ , we may set  $\omega_0 = \sum_1^{m_1} t_j \phi_j$  with the  $t_j$  complex numbers. For  $\omega_0$  having small norm, where the norm is yet to be chosen, we show with the aid of the implicit function theorem that (6.10) has a unique solution  $\omega(\omega_0)$  near  $\omega = 0$ , which then necessarily depends holomorphically on  $\omega_0$ , that is, on the parameters  $t$ . We then require that  $\omega(\omega_0)$  satisfy also (6.9). This requirement gives rise to the "bifurcation equation" which reduces to a finite number of analytic relations on the  $\omega_0$ , that is, on the parameters  $t$ —hence the analytic set in the  $t$  space. The set of solutions thus obtained contains therefore all solutions of (6.9) with small norm.

We must first choose a suitable normed space in which to work so that the implicit function theorem may be applied to (6.9). The appropriate space is furnished by the estimates of the previous section. We may either choose the space of forms with coefficients  $C^{k+\alpha}$ , that is, having derivatives up to order  $k$  which satisfy Hölder conditions with exponent  $\alpha$  (in some fixed coordinate patches covering  $V$ ) as in [27] and [32], or else the space of forms with coefficients in  $H_s$  for  $s$  sufficiently large, the appropriate norms  $\|\cdot\|_{k+\alpha}$  or  $\|\cdot\|_s$  are then easily globally defined and we verify that, say,

$$(6.11) \quad \|[\sigma, \tau]\|_{s-1} \leq \text{constant} \|\sigma\|_s \|\tau\|_s$$

and, in virtue of the estimates of the previous section,

$$(6.12) \quad \|\partial G\phi\|_s \leq \text{constant} \|\phi\|_{s-1}$$

We then easily obtain a unique small solution  $\omega(\omega_0)$  of (6.10) if  $\|\omega_0\|_s$  is small by, say, iterations

$$\omega^{(r+1)} = \omega_0 + \partial G[\omega^{(r)}, \omega^{(r)}], \quad \omega^{(1)} = \omega_0$$

$\omega(\omega_0(t))$  is holomorphic in  $t$ .

We observe the  $(\partial/\partial t_j)\omega(\omega_0(t))|_{t=0} = \phi_j \neq 0$ , that is, the mapping of the tangent space of the  $t$  space at the origin into  $H\Phi^{1,0}$  is injective.

To obtain a solution of (6.9) suppose that  $\omega$  is the solution of (6.10); operating with  $\delta$  we find

$$\delta\omega = \delta\partial G[\omega, \omega] = \square G[\omega, \omega] - \partial \delta G[\omega, \omega]$$

or, by (6.7),

$$(6.13) \quad -\delta\omega + [\omega, \omega] = H[\omega, \omega] + \partial G\bar{\partial}[\omega, \omega]$$

Thus (6.9) is satisfied if and only if the right-hand side is zero; the terms on the right being mutually orthogonal this means that they both vanish:

$$(6.14) \quad H[\omega(\omega_0), \omega(\omega_0)] = 0$$

$$(6.15) \quad \partial G\bar{\partial}[\omega(\omega_0), \omega(\omega_0)] = 0$$

Observe that  $H$  is the projection on the  $m_2$  dimensional space of harmonic elements of  $\Phi^{0,2}$ . Thus equation (6.14) involves a finite number of analytic scalar relations, and these determine a complex analytic set  $M$  in the  $t$ -space. The other relation (6.15) seems to involve an infinite number of relations. It is a consequence of (6.4) to (6.6) however that (6.14) implies (6.15). To see this observe that from (6.4) and (6.5),

$$\partial G\bar{\partial}[\omega, \omega] = 2\partial G[\bar{\partial}\omega, \omega]$$

Substituting the expression for  $\bar{\partial}\omega$  given by (6.13), with  $H[\omega, \omega] = 0$ , we find

$$\begin{aligned} \partial G\bar{\partial}[\omega, \omega] &= 2\partial G[[\omega, \omega], \omega] - 2\partial G[\partial G\bar{\partial}[\omega, \omega], \omega] \\ &= -2\partial G[\partial G\bar{\partial}[\omega, \omega], \omega] \end{aligned}$$

by (6.6). It follows then from (6.11) and (6.12) that

$$\|\partial G\bar{\partial}[\omega, \omega]\|_s \leq \text{constant} \|\partial G\bar{\partial}[\omega, \omega]\|_s \cdot \|\omega\|_s$$

and hence, if  $\|\omega\|_s$  is sufficiently small (which is the case if  $\|\omega_0\|$  is small), it follows that  $\partial G\bar{\partial}[\omega, \omega] = 0$ .

We thus have determined a family  $V_t$  of complex structures (with  $t$  in a complex analytic set) satisfying (1). (In [27] this family was constructed under the assumption that  $H\Phi^{(0,2)}$  vanishes, so that (6.14) and (6.15) then hold automatically for all  $t$ , and  $V_t$  is a full  $m_1$  complex parameter family. For Riemann surfaces,  $n = 1$ , this is trivially the case; in fact the whole problem is linear.)

Kodaira and Spencer [28] have given examples showing that  $V_t$  may be equivalent to  $V_{t'}$  with  $t' \neq t$ .

We have yet to show that  $V_t$  satisfies the conditions of (2), that is, is locally complete. This is proved in [32] but follows most easily from a recent theorem of Kuranishi (to appear).

**Theorem 3.** Let  $\phi$  be a solution of (6.3) with small norm. Then it is "similar" to a solution of (6.9), that is, there is a diffeomorphism  $g$  of  $V$  onto itself which is close to the identity such that the composition  $\phi \circ g$  satisfies  $\partial(\phi \circ g) = 0$ . Furthermore,  $g$  can be made unique as follows:  $g$  can be constructed by exponentiation (via geodesics) of a vector field which is orthogonal to the holomorphic vector fields. In particular if  $\phi(s)$  is a family of solutions of (6.3), the resulting diffeomorphisms  $g(s)$  depend regularly on  $s$ .

This theorem, which is proved in turn with the aid of the implicit function theorem, immediately yields the desired properties of our family  $V_r$ , for if  $\tilde{V}_s$  is a family of complex structure centered at  $V_0$ , the corresponding almost complex structures are described by elements  $\phi(s)$  in  $\Phi^{(0,1)}$  which are solutions of (6.3). According to the theorem there is a diffeomorphism  $g(s)$  such that  $\phi(s) \circ g(s)$  satisfies (6.9) for each  $s$ ; but then it must be a unique one of our  $\omega(\omega_0(t))$  since these contain all solutions near  $\omega = 0$ . Hence there is a well-defined mapping  $t = f(s)$  such that

$$\phi(s) \circ g(s) = \omega(f(s))$$

and this is the desired property.

We recall that for  $t$  small the vector  $(0, 1)$  form  $\omega(t) = \omega(\omega_0(t))$  is holomorphic on the analytic set  $M$ . Suppose now we consider in general a family  $V_t$  as in (1) depending  $C^\infty$  on some complex parameters  $t$  in a neighborhood  $M$  of the origin in some Euclidean  $t$ -space. We say that the resulting family is complex analytic, and that  $M$  has a natural complex analytic structure if we can introduce new coordinates on  $M$  and find a  $C^\infty$  family of diffeomorphisms  $g(t)$  of  $V$  onto itself so that the almost complex structure forms  $\omega(t) \circ g(t)$  depend holomorphically on these new coordinates. In [29] Kodaira and Spencer also take up the question:

3. When can one assert that a given family  $V_t$  is complex analytic?

Kodaira and Spencer proved that certain conditions are sufficient. We shall now describe these. Consider a family  $g(t)$  of diffeomorphisms of  $V$ . If we take the derivative of  $\omega(t)$  with respect to the real or imaginary part of one of the  $t$  variables at  $t = 0$ , we obtain a vector  $(0, 1)$  form  $\phi$  which because of (6.3) satisfies  $\bar{\partial}\phi = 0$ . If, however, we take the derivative of  $\omega(t) \circ g(t)$ , we obtain, as a simple calculation shows,  $\phi + \bar{\partial}\xi$  in place of  $\phi$  where  $\xi$  is a vector

field. On considering all possible families of diffeomorphisms  $g(t)$  we can obtain all possible vector fields  $\xi$ . Thus choosing a vector in the tangent space to  $M$  at 0,  $(T_M)_0$ , determines uniquely an element  $\phi \bmod \partial$  (vector field) with  $\partial\phi = 0$ ; this means that it determines uniquely a harmonic element  $\psi$  (for a unique choice of  $\xi$  makes  $\phi + \partial\xi$  harmonic). Denote by  $\rho_0$  the map  $(T_M)_0 \rightarrow H\Phi^{(0,1)}$ . Clearly for every value of  $t$  there is a similarly well-defined map  $\rho_t: (T_M)_t \rightarrow H_t\Phi_t^{(0,1)}$  (with obvious notation).

The assumptions under which Kodaira and Spencer show that a given family  $V_t$  is complex analytic are, for each  $t$ :

1. The dimension of  $H_t\Phi_t^{(0,1)}$  is independent of  $t$ .
2. The map  $\rho_t$  is injective.
3. The image  $\rho_t((T_M)_t)$  is a complex subspace of  $H_t\Phi_t^{(0,1)}$ .

These assumptions are independent of the particular Hermitian metric introduced on  $V_0$ .

The proof involves construction of an almost complex structure on  $M$  (in a fairly obvious way) and then showing that it satisfies the integrability conditions. The proof of integrability involves further use of elliptic equation theory. A simplified version of the proof has been given in [17].

## 7. $\bar{\partial}$ -NEUMANN PROBLEM

In the previous section, we considered the operator  $\bar{\partial}$  acting on forms on a compact complex analytic manifold and also its adjoint  $\partial$ , which is defined once a Hermitian metric is chosen. In terms of the associated elliptic operator  $\square = \bar{\partial}\partial + \partial\bar{\partial}$ , having a finite dimensional null space, we had the Hodge, Kodaira decomposition theorem. The  $\bar{\partial}$ -Neumann problem is concerned with the same operators acting on forms defined on an open subset  $M$  of  $V$  with compact closure, and satisfying certain boundary conditions (like the Neumann conditions for the classical Laplace equation). We wish to obtain an analogue of the decomposition theorem. For a real manifold with the exterior differentiation operator  $d$  in place of  $\bar{\partial}$  the corresponding problem has been treated by several authors (see [31] for complete references). For that case the boundary conditions satisfy the coerciveness conditions of Section 5. But for  $\bar{\partial}$  the boundary conditions are not coercive and this makes the problem much more difficult and interesting.

The most difficult part of the problem is the proof of regularity at the boundary of "generalized" solutions. If the boundary conditions are not coercive it is known that the lower-order terms may affect the situation (see [21]). We may illustrate this in the case of one complex variable: consider the boundary value problem for a complex valued function  $w$  of a complex variable  $z$  in a disc

$$(7.1) \quad w_{z\bar{z}} + cw = f \quad \text{in } |z| < 1, \quad w_{\bar{z}} = 0 \quad \text{on boundary}$$

with  $c$  a smooth function. (For simplicity we shall suppose that the solution we consider belongs to  $C^2$  in  $|z| \leq 1$  but all remarks are easily extended to generalized, or weak, solutions of the problem.) The boundary condition is not coercive, indeed if  $c \equiv 0$ , the homogeneous problem has an infinite dimensional space of solutions, namely analytic functions. Such solutions need not of course be regular at the boundary. If however  $c \neq 0$  at a boundary point  $z_0$ , it is very easy to prove regularity of the solution at the boundary near  $z_0$  (regularity in the interior is a consequence of the ellipticity of the equation). Dividing the equation by  $c$  and differentiating with respect to  $\bar{z}$ , we obtain the following equation for  $w_{\bar{z}}$ :

$$c^{-1}w_{z\bar{z}\bar{z}} + (c^{-1})_{\bar{z}}w_{z\bar{z}} + w_{\bar{z}} = (c^{-1}f)_{\bar{z}}, \quad w_{\bar{z}} = 0 \quad \text{on boundary}$$

Regarded as a problem for  $w_{\bar{z}}$  the boundary condition  $w_{\bar{z}} = 0$  is of the coercive type (in fact it is the Dirichlet condition), and the well-known local regularity theorems for such imply that  $w_{\bar{z}}$  is regular in some neighborhood of  $z_0$ . Since  $c \neq 0$  near  $z_0$  we have, from (7.1),  $w = c^{-1}(f + w_{\bar{z}})$ , and hence  $w$  is also regular there.

In the  $\bar{\partial}$ -Neumann problem it is the shape of the boundary, the strong pseudoconvexity (in the sense of E. E. Levi), that plays a role, rather than the lower-order terms.

*Strong Pseudoconvexity.* In terms of local analytic coordinates  $z = (z^1, \dots, z^n)$ , suppose that the (smooth) boundary through  $z_0$  is expressed by  $f(z) = 0$  where  $f$  is a smooth real function with  $\text{grad } f(z_0) = 0$ , with  $f < 0$  in  $M$  and  $f > 0$  outside  $M$ . The quadratic form  $f_{z^i\bar{z}^k}a^j\bar{a}^k$  is to be positive definite for all  $a = (a^1, \dots, a^n)$  satisfying  $f_{z^i}a^i = 0$  (using summation convention). It is easily verified that this property is independent of the particular local coordinate or of the function  $f$  representing the boundary.

We now describe the boundary conditions. Observe first that if  $\phi$  and  $\psi$  are, respectively, forms of type  $(p, q)$  and  $(p, q + 1)$  which

are  $C^\infty$  in the closure of  $M$ , then, in terms of the  $L_2$  scalar product,  $(\bar{\partial}\phi, \psi) - (\phi, \bar{\partial}\psi)$  equals a boundary integral which for fixed  $\psi$  vanishes for all such  $\phi$  if and only if  $\nu\psi$  vanishes. At each boundary point the space of forms admits a unique orthogonal decomposition, depending on the metric, into the space of forms having  $\bar{\partial}r$  as a factor and its orthogonal complement; here, for points near the boundary,  $r$  represents the geodesic distance to the boundary measured as positive outside and negative inside  $M$ . In this decomposition we denote the part of the form  $\psi$  having  $\bar{\partial}r$  as a factor by  $\nu\psi$ .

*Boundary Conditions.* The  $\bar{\partial}$ -Neumann boundary conditions for a form  $\phi$  are the following.

$$(7.2) \quad \nu\phi = 0, \quad \nu\bar{\partial}\phi = 0 \quad \text{on boundary } bM$$

We say that  $\phi$  is a harmonic form associated with the  $\bar{\partial}$ -Neumann problem if it satisfies these boundary conditions and is a solution of  $\bar{\partial}\phi = 0$ ,  $\bar{\partial}\phi = 0$ .  $\dot{\mathcal{A}}^{p,q}$  will denote the forms of type  $(p, q)$  which are  $C^\infty$  in the closure of  $M$ ,  $\mathfrak{D}^{p,q}$  the subset satisfying (7.2), and  $\mathfrak{H}^{p,q}$  the smaller subset satisfying also  $\bar{\partial}\phi = \bar{\partial}\phi = 0$ .

Kohn, [30, 31], has recently obtained an analogue of the Hodge-Kodaira decomposition theorem for the  $\bar{\partial}$ -Neumann problem and used it to solve several interesting problems. The work is based on a certain  $L_2$  inequality (7.4) which was proved by Morrey [40] for  $(0, 1)$  forms on small tubular (complex) neighborhoods of real analytic manifolds; Morrey treated the  $\bar{\partial}$ -Neumann problem in such a region in connection with his work on analytic embedding of real analytic manifolds. The inequality is an analogue of the  $L_2$  coercive inequalities for forms  $\phi \in \dot{\mathcal{A}}^{p,q}$ , with  $q > 0$ , in a strongly pseudoconvex domain  $M$  satisfying  $\nu\phi = 0$  on the boundary  $bM$ . In describing it we introduce the following (noninvariant) norm

$$(7.3) \quad E(\phi)^2 = (\phi, \phi) + \int_{bM} |\phi|^2 ds + \|\phi\|_z^2$$

where  $ds$  represents volume element on the boundary, and  $\|\phi\|_z^2$  is defined by means of a fixed covering of the closure of  $M$  by a finite number of holomorphic coordinate patches  $\Omega_j$ , with local coordinates  $z_j$ , as the sum (over  $j$ ) of the integrals in  $\Omega_j$  of the sum of squares of all first derivatives with respect to  $\bar{z}_j^1, \dots, \bar{z}_j^n$  of all components of  $\phi$ .

**Basic Inequality.** For such forms  $\phi$  with  $q > 0$  there is a constant  $C$  such that

$$(7.4) \quad E(\phi)^2 \leq C[(\delta\phi, \delta\phi) + (\partial\phi, \partial\phi) + (\phi, \phi)] = CD(\phi, \phi)$$

We remark that if  $q = 0$ , in which case  $\partial\phi = 0$ , the inequality cannot hold. We can see this immediately if  $p$  also vanishes, for here the inequality would imply that it is possible to estimate  $E(\phi)^2$  in terms of  $(\phi, \phi) = \|\phi\|^2$  for any holomorphic function  $\phi$ —which is clearly not possible.

An important property of the norm  $E(\phi)$  is the fact that it is completely continuous with respect to the  $L_2$  norm, in the sense that if  $\{\phi_j\}$  is a sequence of forms in  $\dot{\mathcal{K}}^{p,q}$  having uniformly bounded norms  $E(\phi_j)$ , a subsequence converges in the  $L_2$  norm  $\|\cdot\|$ .

Before describing Kohn's result let us pursue a natural line of reasoning suggested by the inequality (7.4) leading to a decomposition theorem in  $L_2$ . This is to form the completion,  $\mathcal{K} = \mathcal{K}^{p,q}$ , in the norm  $D(\phi, \phi)$  of forms in  $\dot{\mathcal{K}}^{p,q}$  satisfying  $\nu\phi = 0$  on  $bM$ , and to define a "generalized" solution  $\phi$  in  $\mathcal{K}$  of the boundary value problem

$$(7.5) \quad \square\phi + a\phi = \alpha, \quad \nu\phi = 0, \quad \nu\delta\phi = 0 \text{ on } bM$$

given  $a \in C^\infty$  in the closure of  $M$ , as an element  $\phi \in \mathcal{K}^{p,q}$  satisfying

$$(7.6) \quad (\delta\phi, \delta\psi) + (\partial\phi, \partial\psi) + (a\phi, \psi) = (\alpha, \psi) \text{ for all } \psi \in \mathcal{K}^{p,q}$$

Observe that the boundary condition  $\nu\delta\phi = 0$  is obtained as a "natural" boundary condition; if  $\phi$  is regular in  $M + bM$  then (7.5) and (7.6) are equivalent. In particular the space  $\mathcal{K}^{p,q}$  of generalized harmonic solutions of the  $\bar{\partial}$ -Neumann problem is defined as those functions in  $\mathcal{K}^{p,q}$ , satisfying

$$(\delta\phi, \delta\psi) + (\partial\phi, \partial\psi) = 0 \quad \text{for all } \psi \in \mathcal{K}^{p,q}$$

As a consequence of the basic inequality (7.4) and the complete continuity of the norm  $E(\phi)$ , we see that for  $q > 0$  the spaces  $\mathcal{K}^{p,q}$ , and a fortiori also  $\dot{\mathcal{K}}^{p,q}$ , are finite dimensional. More generally if  $q > 0$  the space of solutions of the homogeneous equation (7.6),  $\alpha = 0$ , is finite dimensional. The space  $\mathcal{K}^{p,0}$  is in general infinite dimensional. Another consequence is the fact that, for  $q > 0$ , the set of  $(p, q)$  forms  $\alpha \in L_2$  for which (7.5) is solvable in the generalized sense (7.6) is a closed subspace of  $L_2$ , the range of the "generalized" operator  $\square + a$ . This is derived in a straightforward way

using the closed graph theorem. Kohn proves in addition by a rather simple argument that the same is true also for  $q = 0$  despite the lack of complete continuity in this case.

Consider now elements  $\psi \in \mathcal{K}^{p,q} \ominus \mathcal{K}^{p,q}$ ,  $q > 0$ , that is, orthogonal (in  $L_2$ ) to  $\mathcal{K}^{p,q}$ . Then there is a constant  $C_1$  such that

$$(7.4)' \quad \|\psi\|^2 \leq C_1[(\delta\psi, \delta\psi) + (\partial\psi, \partial\psi)] \quad \text{for all } \psi \in \mathcal{K}^{p,q} \ominus \mathcal{K}^{p,q}$$

(this follows easily by an indirect argument using the complete continuity of  $E(\phi)$ ). We see that if  $\phi$  is a solution of (7.6) with  $\alpha = 0$  then  $(\alpha, \psi)$  vanishes for  $\psi \in \mathcal{K}^{p,q}$ , that is,  $\alpha \in L_2 \ominus \mathcal{K}$ . From (7.4)' we see conversely, with the aid of the representation theorem for linear functionals in Hilbert space, that for any  $\alpha \in L_2 \ominus \mathcal{K}^{p,q}$ ,  $q > 0$ , the equation (7.6) is solvable and there is a unique solution in  $\mathcal{K}^{p,q} \ominus \mathcal{K}^{p,q}$ . Denote this by  $N\alpha$  and define  $N\alpha$  as zero for  $\alpha \in \mathcal{K}^{p,q}$  and extend by linearity to all  $\alpha$  in  $L_2$ . Thus if  $H\alpha$  denotes the projection (in  $L_2$ ) of any  $(p, q)$  form  $\alpha$  into  $\mathcal{K}^{p,q}$ , we obtain the following decomposition, for  $q > 0$ ,

$$(7.7) \quad \alpha = H\alpha + \square N\alpha$$

where  $\square\phi$  is understood in the generalized sense as the element  $\alpha$  such that (7.6) holds.

This is the  $L_2$  decomposition theorem for  $(p, q)$  forms,  $q > 0$ ;  $N$  here is the Neumann function, analogous to the Green's function.

The main part of Kohn's work, and the most difficult, is the proof that for  $q > 0$  the operator  $N$  maps  $\dot{\mathcal{K}}^{p,q}$  into  $\dot{\mathcal{K}}^{p,q}$ , that is, that the generalized solutions of (7.6) for  $\alpha \in \dot{\mathcal{K}}^{p,q}$  (i.e., of class  $C^\infty$  in the closure of  $M$ ) are themselves in  $\dot{\mathcal{K}}^{p,q}$ ,  $q > 0$ . This is proved by a local analysis at the boundary using techniques developed by Hörmander [25] in connection with his treatment of regularity for the Cauchy problem (as Theorem 2, Section 2, Lecture 1). This involves proving first the regularity of tangential derivatives. The proof is too intricate to be described here.\*

\* Simplified proofs of the regularity at the boundary have been obtained by C. B. Morrey, and more recently by J. J. Kohn and L. Nirenberg. L. Hörmander has given a modified approach to the  $\bar{\partial}$ -Neumann problem (to appear) which avoids the question of regularity at the boundary; it uses a weaker regularity theorem due to P. Lax and R. Phillips [34]. Although the decomposition theorem he obtains is not as strong as Kohn's, it is sufficient for many applications to functions of several complex variables.

Having the differentiability, we easily verify that

$$(7.8) \quad \square N = N \square, \quad \partial N = N \partial, \quad \text{and} \quad \bar{\partial} N = N \bar{\partial}$$

in the sense that  $(\partial N \alpha, \beta) = (\partial \alpha, N \beta)$  for  $\alpha, \beta \in \mathcal{K}^{p,q}$ . Furthermore,  $N$  is a compact operator of  $L_2$  into  $L_2$  for  $q > 0$ .

Observe that

$$(7.9) \quad H \alpha = \alpha - \partial N \bar{\partial} \alpha \quad \text{for } \alpha \in \mathcal{K}^{p,0}$$

Kohn derives an analogous decomposition theorem, and operator  $N$  (which is not compact, however), also in case  $q = 0$ . Some additional argument is required.

At this point we would like to indicate how the main inequality (7.4) is proved. Since the operator  $\square$  is elliptic the estimate of all first derivatives in any compact subset in terms of  $D(\phi, \phi)$  follows from standard elliptic theory. Using a partition of unity we reduce the problem to that of establishing (7.4) for functions with support in some coordinate patch near a boundary point. For such a function, (7.4) is derived by a careful use of Green's theorem. We shall carry it out for the very special situation of  $(0, 1)$  forms in a strongly pseudoconvex domain in  $C^n$ . For such a form  $\phi = \phi_{j,k} d\bar{z}^j$  we have, using  $\bar{\partial}$  and  $\partial$  to denote  $\partial/\partial \bar{z}^j$  and  $\partial/\partial z^j$ ,

$$(7.10) \quad \begin{aligned} \|\bar{\partial} \phi\|^2 + \|\partial \phi\|^2 &= \frac{1}{2} \int_M \sum_{j \neq k} |\phi_{j,k} - \phi_{k,j}|^2 + \int_M \sum_{j,k} \phi_{j,j} \overline{\phi_{k,k}} \\ &= \int_M \sum_{j,k} |\phi_{j,k}|^2 + \int_M \sum_{j,k} (\phi_{j,j} \overline{\phi_{k,k}} - \phi_{j,k} \overline{\phi_{k,j}}) \end{aligned}$$

The first integral on the right of (7.10) is  $\|\phi\|^2$  whereas the integrand in the second is a divergence expression and so is equal to a boundary integral

$$(7.11) \quad \sum_{j,k} \int_{\partial M} \overline{\phi_{k,k}} (\phi_{j,j} n_k - \phi_{j,k} n_j) ds$$

where  $n_k, n_{\bar{k}}$  are the (complex) components of the normal in the sense that  $n_k$  and  $n_{\bar{k}}$  are proportional (with a positive factor  $b$ ) to  $f_{,k}, f_{,\bar{k}}$  where  $f$  is the function representing the boundary locally by  $f = 0$ .

We now have to use the boundary condition  $n\phi = 0$  which in this case is simply  $\sum n_k \overline{\phi_{k,k}} = 0$  or  $\sum f_{,k} \overline{\phi_{k,k}} = 0$ . On taking complex conjugate this means  $\sum n_{\bar{k}} \phi_{k,k}$  so that the first term in (7.11) drops

out. We may wonder how the second term can be estimated in terms of the square integral of  $|\phi|$  on the boundary, since it involves first derivatives of  $\phi$ . The reason is that it may be expressed by a *tangential* derivative of  $\nu\phi$  modulo a quadratic in  $\phi$ . In fact on the boundary we have (using summation convention)  $f_{,\bar{k}}\bar{\phi}_{\bar{k}} = 0$ . This means that the operator  $\bar{\phi}_{\bar{k}}(\partial/\partial\bar{z}^k)$  operates tangentially on the boundary. Hence in particular we find on applying it to  $f_{,j}\phi_j = 0$ ,

$$\bar{\phi}_{\bar{k}} \frac{\partial}{\partial\bar{z}^k} (f_{,j}\phi_j) = 0 \quad \text{on } bM$$

or

$$-\bar{\phi}_{\bar{k}} f_{,j} \phi_{j,\bar{k}} = \bar{\phi}_{\bar{k}} f_{,j\bar{k}} \phi_j$$

Thus we find that the integral in (7.11) equals

$$\int_{bM} b f_{,j\bar{k}} \phi_j \bar{\phi}_{\bar{k}} ds$$

and according to strong pseudoconvexity the integral is a positive definite quadratic form. This completes the derivation of the inequality.

The reason that pseudoconvexity does not play a role in the analogous problem on real manifolds with  $d$  in place of  $\delta$  is that (7.4) holds on any domain  $M$  (even without pseudoconvexity) with, in fact  $\|\phi\|_z^2$  replaced by the integral of the sums of squares of *all* first derivatives of the components of  $\phi$ , and hence we do not need the boundary integral analogous to (7.11) to be positive. We note, however, that if we consider 1-forms  $\Sigma \phi_j dx^j$  in a strong pseudoconvex domain in  $R^n$  (which is simply equivalent to strict convexity), the boundary integral is indeed positive, and we find, assuming  $n\phi = 0$  on  $bM$

$$\frac{1}{2} \|d\phi\|^2 + \|\delta\phi\|^2 = \sum_{j,k} \int_M \left| \frac{\partial \phi_j}{\partial x^k} \right|^2 + \int_{bM} R^{jk} \phi_j \bar{\phi}_k ds$$

where the quadratic form in the boundary integral is positive definite.

Kohn makes a number of applications of the decomposition theorem. In particular, he gives a new proof, along the lines suggested by Spencer, of the fact that an almost complex structure satisfying the integrability conditions corresponds to a complex

structure. Although the proof is perhaps more involved than the one given in [41], it is interesting in that it works completely in the framework of linear theory in contrast to the proof in [41].

The idea of the proof is the following. Consider an almost complex structure given in some neighborhood of the origin in  $C^n$  by forms as in Section 6:

$$(7.12) \quad \phi^j = dz^j + \omega_{\bar{k}}^j dz^{\bar{k}}, \quad j = 1, \dots, n$$

We may choose these so that the  $\omega_{\bar{k}}^j$  vanish at the origin. Using these, one may again decompose the space of forms into forms of associated type  $(p, q)$ , namely forms spanned by exterior products of  $p$  of these and  $q$  of the complex conjugate forms,  $d\bar{z}^j + \overline{\omega_{\bar{k}}^j} dz^{\bar{k}}$ . Furthermore, we may decompose  $d$  into  $\bar{\partial} + \partial$ , using the unique decomposition  $df = a_j \phi^j + b_{\bar{j}} \bar{\phi}^{\bar{j}}$  by  $\partial f = a_j \phi^j$ ,  $\bar{\partial} f = b_{\bar{j}} \bar{\phi}^{\bar{j}}$ . The integrability condition now takes the elegant form  $\bar{\partial}^2 = 0$ .

Kohn remarks that the foregoing estimates and the decomposition theorem may be derived for such operators  $\bar{\partial}$ .

Now replace the forms  $\phi^j$  above by the forms  $\phi_t^j$  where  $\omega_{\bar{k}}^j(z, \bar{z})$  is replaced by  $\omega_{\bar{k}}^j(tz, t\bar{z})$ ,  $t > 0$ . We verify easily that the  $\phi_t^j$  still satisfy the integrability conditions. Let  $\bar{\partial}_t$  be the corresponding operator; for  $t = 0$  it corresponds to the  $\bar{\partial}$  of the underlying complex structure in  $C^n$ . If we take for  $M$  the unit ball, it is certainly strongly pseudoconvex for  $\bar{\partial}_0$  and hence also for  $\bar{\partial}_t$  for small  $t$ . Let  $N_t$  be the corresponding Neumann operator in the decomposition theorem and apply the decomposition (7.7) to the function  $z^j$ . According to (7.9),

$$z_t^j = z^j - \partial_t N_t \bar{\partial}_t z^j$$

satisfies  $\bar{\partial}_t z_t^j = 0$ .

If for small  $t$  the derivatives of the term  $\beta_t = \partial_t N_t \bar{\partial}_t z^j$  are small at the origin then the derivatives of  $z_t^j$  are close to those of  $z^j$  there, and hence the  $z_t^j$  may be introduced as new coordinates in some neighborhood of the origin. The functions  $\zeta^j = z_t^j \left( \frac{z}{t}, \frac{\bar{z}}{t} \right)$  are then the desired holomorphic coordinates near the origin for the original almost complex structure given by (7.12).

To estimate the derivatives of  $\beta_t$ , observe first that

$$\|\beta_t\| \leq \|\partial_t N_t\| \|(\bar{\partial}_t - \bar{\partial}_0)z^j\|$$

where  $\|\partial_t N_t\|$  denotes the norm of the operator  $\partial_t N_t$  as a mapping of the space of square integrable forms into itself. Since the spaces

$\mathcal{H}^{p,q}$ ,  $q > 0$ , for  $\delta_0$  are null, the same is true for  $\delta_t$  for small  $t$  (because the dimension of null spaces of elliptic operators depending on a parameter is upper semicontinuous in the parameter). Consequently we may show that the operator  $N_t$  depends smoothly in  $t$  and hence the norms  $\|\partial_t N_t\|$  are uniformly bounded for small  $t$ . Since  $\delta_t - \delta$  has small coefficients we see finally that  $\|\beta_t\|$  is small.

It follows from estimates for elliptic equations applied to the equation  $\bar{\partial}_t \beta_t = (\delta_t - \delta)z^j$  that then all derivatives of  $\beta_t$  are small near the origin for small  $t$ —the desired result.

## 8. INDEX OF ELLIPTIC OPERATORS

We have observed in Section 5 that the index for an elliptic differential operator with coercive boundary conditions is finite. Recently there has been considerable interest in calculating the index for general problems (see Gelfand [18]). Previously the index had been evaluated for a variety of first- and second-order equations in the plane—starting with the so-called Riemann-Hilbert problem which is essentially the following: to determine a function  $u$  in the unit disc satisfying

$$\Delta u = f, \quad \nu \operatorname{grad} u = \phi \text{ on boundary}$$

where  $\nu$  is a unit vector defined on the boundary as a function of arc length. The index of this problem is equal to  $2(1 - n)$  where  $n$  is the rotation number of  $\nu$ , going around the circle counterclockwise. Problems in the plane are treated in great detail in the book [54] by Vekua, which also contains an extensive bibliography.

For higher dimensions several authors, Agranovic, Dynin, Volpert, Seeley, Cordes, Koppelman, and others, have obtained results in special cases. We refer to [3] and [12] for other references. Atiyah and Singer [5] have determined a general formula for the index of an elliptic operator on any compact oriented manifold. Applying it to various particular differential operators arising geometrically, they obtain in a unified manner many known topological results and also new results—for example, the Hirzebruch-Riemann-Roch theorem for compact manifolds, which had previously been proved only for projective algebraic manifolds. The formula for the index is expressed in terms of certain topological and analytic invariants associated with the manifold and the oper-

ator. It is unfortunately too complicated to be explained here; we can merely make a few remarks about some of the analytic techniques that are involved.

Let us first consider some special cases treated by Atiyah and Singer. Consider a compact Riemannian manifold  $V$  of dimension  $n = 2l$ ; let  $\delta$  denote the formal adjoint of the exterior differential operator  $d$ . On the space of all complex valued forms the operator  $d + \delta$  is elliptic, but also self-adjoint so its index is zero. But by considering  $d + \delta$  on a subspace it is possible to get index  $\neq 0$ .

1. If  $*$  is the usual isomorphism of  $p$  forms to  $n - p$  forms defined with the aid of the metric then  $(*)^2 = (-1)^k$ . Let  $X_1(X_2)$  denote the space of forms with even (odd) degree. Then  $d + \delta$  maps  $X_1$  into  $X_2$  and restricted to  $X_1$  it is still elliptic, but no longer self-adjoint. Applying their formula for the index they obtain the following, which also follows easily from the Hodge theory and the Gauss-Bonnet theorem,

$$\sum_{p=0}^{2l} (-1)^p \dim H^p = \chi(V), \quad \text{the Euler characteristic}$$

Here  $H^p$  is the space of harmonic  $p$  forms.

2. By decomposing the space of forms in a different way into two spaces  $X_1 + X_2$  with  $d + \delta: X_1 \rightarrow X_2$ , and calculating the index of  $d + \delta$ , they obtain the so-called Hirzebruch index theorem.

3. If  $V$  is a complex manifold with Hermitian metric and the operator  $\bar{\partial} + \partial$  is applied to functions with values in a certain holomorphic vector bundle, they obtain the Hirzebruch-Riemann-Roch theorem.

Turning now to the analytic techniques we remark that by introducing suitable normed topologies it is possible to represent the elliptic operator as a bounded operator. Then since the index of an operator is unchanged under slight perturbation or by addition of a compact operator, it follows that the lower-order terms of the operator, which act as a compact operator in comparison with the highest-order terms, do not affect the index. Thus the index of the operator depends only on the highest-order terms. Furthermore, if the operator is self-adjoint, the index is obviously zero. We mention also that the index of the product of two operators is the sum of their indices.

The elliptic operators that they treat are for systems of functions,

and since these are to include forms, which cannot be expressed globally as a system of functions over the manifold, they consider elliptic operators operating on smooth sections of a complex vector bundle  $E$  over the manifold  $V$  and mapping into sections of another complex vector bundle  $F$  over  $V$  with the fibers having the same dimension  $N$ . Since the index of an elliptic operator is invariant under deformation of the operator, they seek to deform the operator to one composed of fairly standard ones. But it seems to be a difficult problem to determine the connected components of elliptic operators; instead they consider a more general class of operators including the elliptic ones as a very special case: singular integral operators whose symbol (represented locally as an  $N \times N$  matrix defined on the cotangent bundle) does not vanish (i.e., has non-zero determinant). Singular integral operators have been extended by Seeley [50] to compact manifolds (for the scalar case, but it is also easily done in the vector bundle framework), and they have finite index if the symbol does not vanish (see Section 4). Furthermore, the index depends only on the symbol. The formula obtained in [5] is valid, more generally, for such "elliptic" singular integral operators.

With suitable introduction of an equivalence on the set of all elliptic symbols (with the property that equivalent symbols have the same index), the equivalence classes form an abelian semigroup, and they reduce the calculation of the index to that for a very special elliptic operator, namely the operator  $d + \delta$  acting in the decomposed spaces of example 2, tensored, however, with a suitable vector bundle  $W$ . To calculate the index for this they follow Hirzebruch, who carried it out for  $W = 1$ . A basic step is the proof that if  $V$  (of even dimension) is the boundary of a manifold  $Y$  with a vector bundle  $U$  on it, which at the boundary reduces to  $W$ , the index of the particular operator is zero. To do this they make use of an extension of the operator  $d + \delta$  (with the given decomposition  $X_1 + X_2$ ) into  $Y$ , and show that the resulting boundary value problems, prescribing the  $X_1$  component on the boundary, are the coercive type mentioned in Section 5, and that the  $X_2$  boundary component of the solution may be expressed by means of a singular integral operator  $T$  in terms of the given  $X_1$  boundary component; and the index of  $T$  is then shown to be zero.

We cannot go into any detail here since the proofs use a great deal of topology as well as analysis.

## REFERENCES

1. S. Agmon, "The coerciveness problem for integro-differential forms," *J. Analyse Math.* **6** (1958) 183-223.
2. S. Agmon, A. Douglis, and L. Nirenberg, "Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions," I, II, *Comm. Pure Appl. Math.* **12** (1959) 623-727; **17** (1964) 35-92.
3. M. S. Agranovic and A. D. Dynin, "General boundary value problems for elliptic systems in higher dimensional regions," *Dokl. Akad. Nauk. SSSR* **146** (1962) 511-514. *Soviet Mathematics* **3** (1962), 1323-1327.
4. L. V. Ahlfors, "Teichmüller spaces," *Proc. Intern. Congr. Math. Stockholm* (1962).
5. M. F. Atiyah and I. M. Singer, "The index of elliptic operators on compact manifolds" (mimeographed).
6. L. Bers, "Spaces of Riemann surfaces," *Proc. Intern. Congr. Math. Edinburgh* (1958) 14-21.
7. F. Browder, "A priori estimates for solutions of elliptic boundary value problems," I, II, *Indag. Math.* **22** (1960) 145-159, 160-169.
8. A. P. Calderon, "Uniqueness in the Cauchy problem for partial differential equations," *Amer. J. Math.* **80** (1958) 16-36.
9. A. P. Calderon, "Integrales singulares y sus aplicaciones a ecuaciones diferenciales hiperbolicas," Univ. Buenos Aires, Fac. Ciencias Exactas y Nat. Dep. de Matem. (1960).
10. A. P. Calderon and A. Zygmund, "Singular integral operators and differential equations," *Amer. J. Math.* **79** (1957) 901-921.
11. P. Cohen, "The non-uniqueness of the Cauchy problem," Technical Report No. 93, Stanford University (1960).
12. H. O. Cordes, Lecture notes at University of California, Berkeley (1962).
13. A. Douglis and L. Nirenberg, "Interior estimates for elliptic systems of partial differential equations," *Comm. Pure Appl. Math.* **8** (1955) 503-538.
14. D. G. Figueiredo, "The coerciveness problem for forms over vector valued functions," *Comm. Pure Appl. Math.* **16** (1963).
15. A. Friedman, *Generalized functions and partial differential equations*, Englewood Cliffs, N.J. (1963) Prentice-Hall.
16. K. O. Friedrichs, "Symmetric positive linear differential equations," *Comm. Pure Appl. Math.* **11** (1958) 333-418.
17. A. Frölicher, E. T. Kobayashi, and A. Nijenhuis, "Deformation theory of complex manifolds," Technical Report No. 10, 1959, University of Washington.
18. I. M. Gelfand, "On elliptic equations," *Uspekhi Mat. Nauk* **15** (1960) 121; *Russian Math. Surveys* **15** (1960) No. 3, 113-123.
19. I. M. Gelfand and G. E. Silov, *Generalized functions: Vol. 1, Generalized functions and operations on them; Vol. 2, Spaces of fundamental functions and generalized functions; Vol. 3, Some questions in the theory of differential equations*, Moscow (1958-1959).

20. L. Hörmander, "On the theory of general partial differential operators," *Acta Math.* **94** (1955) 161-248.
21. L. Hörmander, "On the regularity of the solutions of boundary problems," *Acta Math.* **99** (1958) 225-264.
22. L. Hörmander, "On the uniqueness of the Cauchy problem," I, II, *Math. Scand.* **6** (1958) 213-225; **7** (1959) 177-190.
23. L. Hörmander, "Differential operators of principal type. Differential equations without solutions," *Math. Ann.* **140** (1960) 124-146 and 169-173.
24. Hörmander, *On the range of differential and convolution operators*, Institute for Advanced Study, Princeton, (mimeographed) April 1961.
25. L. Hörmander, *Linear partial differential operators*, Berlin (1963) Springer.
26. F. John, "Continuous dependence on data for solutions of partial differential equations with a prescribed bound," *Comm. Pure Appl. Math.* **13** (1960) 551-585.
27. K. Kodaira, L. Nirenberg, and D. C. Spencer, "On the existence of deformations of complex analytic structures," *Ann. of Math.* **68** (1958) 450-459.
28. K. Kodaira and D. C. Spencer, "On deformations of complex analytic structures," I, II, *Ann. of Math.* **67** (1958) 328-466; III, Stability theorems for complex structures, *Ann. of Math.* **71** (1960) 43-76.
29. K. Kodaira and D. C. Spencer, "Existence of complex structure on a differentiable family of deformations of compact complex manifolds," *Ann. of Math.* **70** (1959) 145-166.
30. J. J. Kohn, "Solution of the  $\bar{\partial}$ -Neumann problem on strongly pseudoconvex manifolds. Regularity at the boundary of the  $\bar{\partial}$ -Neumann problem," *Proc. Nat. Acad. Sci.* **47** (1961) 1198-1202; **49** (1963) 206-213.
31. J. J. Kohn, "Harmonic integrals on strongly pseudoconvex manifolds," I, II, *Ann. of Math.* (to appear).
32. M. Kuranishi, "On the locally complete families of complex analytic structures," *Ann. of Math.* **75** (1962) 536-577.
33. P. D. Lax, "The  $L_2$  operator calculus of Mihlin, Calderon and Zygmund" (mimeographed lecture), New York University (1963).
34. P. D. Lax and R. S. Phillips, "Local boundary conditions for dissipative symmetric linear differential operators," *Comm. Pure Appl. Math.* **13** (1960) 427-455.
35. H. Lewy, "An example of a smooth linear partial differential equation without solution," *Ann. of Math.* (2) **66** (1957) 155-158.
36. J. L. Lions and E. Magenes, "Problemi ai limiti non omogenei. Problèmes aux limites non homogènes," I, III, IV, V, *Ann. Scuola Norm. Sup. Pisa* **14** (1960) 269-308; **15** (1961) 39-101; **15** (1961) 311-326; **16** (1962) 1-44; II, *Ann. Inst. Fourier* **11** (1961) 137-178.
37. B. Malgrange, "Existence et approximation des solutions des équations aux dérivées partielles et des équations de convolution," *Ann. Inst. Fourier, Grenoble* **6** (1955-1956) 271-355.
38. B. Malgrange, "Sur la propagation de la régularité des solutions des équations à coefficients constants," *Bull. Math. Soc. Math. Phys., Roumanie*, to appear.
39. S. G. Mihlin, "Singular integral equations," *Uspekhi Mat. Nauk USSR* **NS 3** (25) (1948) 29-112.

40. C. B. Morrey, "The analytic embedding of abstract real manifolds," *Ann. of Math.* **68** (1958) 159-201.
41. A. Newlander and L. Nirenberg, "Complex analytic coordinates in almost complex manifolds," *Ann. of Math.* **65** (1957) 391-404.
42. A. Nijenhuis and W. B. Woolf, "Some integration problems in almost-complex and complex manifolds," *Ann. of Math.*, **77** (1963) 424-489.
43. L. Nirenberg, *A complex Frobenius theorem*, Seminars on Analytic Functions, Institute for Advanced Study, Princeton (1957) 172-189.
44. L. Nirenberg, "Functional analysis," Lecture notes, New York University, *Courant Institute of Mathematical Sciences* (1961).
45. L. Nirenberg, "Some aspects of linear and nonlinear partial differential equations," *Proc. Intern. Congr. Math., Stockholm* (1962).
46. L. Nirenberg and F. Trèves, "Solvability of a first order linear partial differential equation," *Comm. Pure Appl. Math.*, **16** (1963) 331-351.
47. J. Peetre, "Elliptic partial differential equations of higher order," Lecture notes, University of Maryland, Institute for Fluid Dynamics and Applied Mathematics, 1962.
48. L. Sarason, "On weak and strong solutions of boundary value problems," *Comm. Pure Appl. Math.* **15** (1962) 237-288.
49. M. Schechter, "On  $L_p$  estimates and regularity I," *Amer. J. Math.*, **85** (1963) 1-13.
50. R. T. Seeley, "Singular integral operators on compact manifolds," *Amer. J. Math.* **81** (1959) 658-690.
51. Séminaire Schwartz, *Unité du problème de Cauchy*, University of Paris (1959-1960).
52. F. Trèves, "Relations de domination entre opérateurs différentiels," *Acta Math.* **101** (1959) 1-139.
53. F. Trèves, "Lectures on linear partial differential equations with constant coefficients," Lecture Notes No. 27, Inst. de Mat. Pura e Aplicada de Conselho Nacional de Pesquisas, Rio de Janeiro (1961).
54. I. N. Vekua, *Generalized analytic functions*, Fizmatgiz. Moscow (1959). English translation New York (1962) Pergamon Press.
55. M. Zerner, "Solutions de l'équation des ondes présentant des singularités sur une droite," *C. R. Acad. Sci. Paris* **250** (1960) 2980-2982.

# 2

## Generators and Relations in Groups —The Burnside Problem

*Marshall Hall, Jr.*

### 1. INTRODUCTION

Many of the natural problems on generators and relations in groups are difficult and some are impossible. Given a group  $G$  generated by a finite number of elements subject to a finite number of relations, the problem of deciding whether or not a particular word in the generators is the identity is known as the "word problem." The word problem is impossible in that it is recursively unsolvable, as was shown first by Novikov in 1955. From this it follows that a host of other reasonable sounding problems are also unsolvable.

The unsolvability of these problems in their most general form invites us to the consideration of special cases where we may solve the problems. The most natural way to consider problems on generators and relations in groups is to consider the groups as presented as factor groups of free groups. Sections 2 through 6 give a brief discussion of some of the main results. Section 6 gives a sketch of a method of Tartakovskii, which yields a solution of the word problem under suitable restrictions.

The monograph *Generators and Relations for Discrete Groups*, by Coxeter and Moser [8] contains a wealth of material and a large number of particular examples which the reader will find most

rewarding. It has been possible to include only a very little of this here.

The particular problem which is treated here with some degree of thoroughness is the Burnside problem. The present formulation of this problem is the question, "If  $G$  is a group generated by a finite number of elements with the property that there is a positive integer  $n$  such that  $x^n = 1$  for every  $x$  in  $G$ , is  $G$  necessarily finite?" In his original paper Burnside (7) formulated this as a question and not as a conjecture, although some writers speak of "Burnside's conjecture." The groups  $G$  are known to be finite if  $n = 2, 3, 4$ , or 6, and Novikov announced in 1959 that the groups are infinite for  $n \geq 72$ . But Novikov's proof has not yet appeared, and it may be that since his announcement he has found a gap in his proof. Sections 6, 7, and 8 deal with the Burnside problem. Section 6 deals with the elementary and direct methods applicable to  $n = 2, 3, 4$ . Section 7 discusses the application of the theory of  $p$ -length developed by P. Hall and G. Higman [19]. Section 8 describes the use of commutators and associated Lie rings in attacking the restricted Burnside problem and in particular Kostrikin's solution of the restricted Burnside Problem for prime exponents.

## 2. PRESENTATION OF GROUPS

Suppose a group  $G$  is generated by elements  $x$  and  $y$  satisfying relations  $x^2 = y^2 = (xy)^3 = 1$ . Here  $G$  is finite of order 6. The distinct elements of  $G$  are 1,  $x$ ,  $y$ ,  $xy$ ,  $yx$ ,  $xyx$ . In this simple case let us consider the meaning of the statements and the questions they raise.

1. " $G$  is generated by elements  $x$  and  $y$ ." This means that  $x$  and  $y$  are elements of  $G$ , and that every element of  $G$  has at least one expression as a finite product of the elements  $x$ ,  $x^{-1}$ ,  $y$ , and  $y^{-1}$ .

2. "The distinct elements of  $G$  are 1,  $x$ ,  $y$ ,  $xy$ ,  $yx$ ,  $xyx$ ." This means that every element of  $G$  can be put into one of these six forms as a consequence of the relations  $x^2 = y^2 = (xy)^3 = 1$ , and that there is no self-contradiction in these relations or any further equalities consequences of these relations.

Since  $x^2 = 1$ , then  $x^{-1}(x^2) = x^{-1}$  or  $(x^{-1}x)x = x^{-1}$  or  $x = x^{-1}$ . Similarly from  $y^2 = 1$  follows  $y = y^{-1}$ . Hence a finite product of  $x$ ,  $x^{-1}$ ,  $y$ ,  $y^{-1}$  can be written as a finite product of  $x$  and  $y$ . In a finite product  $g = a_1a_2 \cdots a_{i-1}a_ia_{i+1}a_{i+2} \cdots a_t$ , where for

$j = 1, \dots, t$  either  $a_j = x$  or  $a_j = y$ , if two consecutive  $a$ 's are equal say  $a_i = a_{i+1} = x$  or  $a_i = a_{i+1} = y$ , since  $x^2 = y^2 = 1$ , we have  $a_i a_{i+1} = 1$ , and we may rewrite  $g = a_1 \cdots a_{i-1} a_{i+2} \cdots a_t$ . Thus  $g$  is either the identity or an alternating product of  $x$ 's and  $y$ 's. Also from  $(xy)^3 = 1$  we find  $xyxyxy = 1$ ,  $xyx = x^{-1}y^{-1}x^{-1} = xyx$ . From this we see that every  $g$  can be put in one of the forms  $1, x, y, xy, yx, xyx$ , and so  $G$  has at most six elements. But if we consider the permutations  $X = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ ,  $Y = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$ , we find that  $X^2 = Y^2 = (XY)^3 = 1$  and these permutations generate the symmetric group on three letters, and in this the six elements  $1, X, Y, XY, YX, XYX$  are all different. Here mapping  $x$  into  $X$  and  $y$  into  $Y$  we have what we call a realization of the group  $G$ . This realization shows that the relations are not self-contradictory, and that no further equalities are consequences of the relations.

We may easily dispose of one natural question which arises here. When are defining relations for a group  $G$  consistent? The answer is that *any* set of defining relations is consistent. For if  $G$  is generated by elements  $x_i$  and relations  $f_j(x) = 1$ , then, if every  $x_i = 1$ , every relation  $f_j(x) = 1$  is satisfied. Here  $G$  is the identity and is a realization of all the relations. The group  $G$  consisting of the identity alone is thus a realization of every group given by generators and relations, and this we call the trivial realization. If the only realization of a group  $G$  with generators and relations is the trivial one we say that  $G$  is trivial. We have already reached an unsolvable problem. No effective procedure exists to determine whether or not a group  $G$  given by generators and relations is trivial. The group  $G$  generated by the three elements  $x, y, z$  satisfying relations

$$(2.1) \quad x^{-1}yx = y^2, y^{-1}zy = z^2, z^{-1}xz = x^2$$

is trivial, but the reader may find that it is not "trivial" to prove this.

What may be said about a group  $G$  with generators but no relations? It is conceivable that in every group with say  $r$  generators there is some nontrivial combination which is the identity, perhaps the same one for all such groups, perhaps a number of relations exist such that at least one is satisfied in any group with  $r$  generators. Groups with no relations exist and are called free groups. This relatively simple state of affairs is not obvious. In a free Lie ring

there are many nontrivial relations [12], and the writer knows of no satisfactory definition of a free field.

Let  $I$  be an index system. For the most part we shall be interested in the case in which there are only a finite number of indices and  $I$  is the set  $1, 2, \dots, n$ . For an infinite set  $I$  of indices we shall assume the indices to be well ordered. We define a set  $S$  of elements  $s_i = s_i^{-1}$ ,  $i \in I$  and their formal inverses  $s_i^{-1}$ . A word or string is either void (written 1) or a finite succession  $a_1 a_2 \cdots a_t$  where each  $a_i$  is one of the  $s_j^\epsilon$ ,  $\epsilon = \pm 1$ . A word  $w$  is a *reduced word* if it is void or if  $w = a_1 a_2 \cdots a_t$  and no pair  $a_i a_{i+1}$ ,  $i = 1, \dots, t-1$  is of the form  $s_j^\epsilon s_j^{-\epsilon}$ ,  $\epsilon = \pm 1$ . Here  $t$  is called the *length* of  $w$ . Two words  $f_1$  and  $f_2$  are *adjacent* if they are of the form  $f_1 = g s_j^\epsilon s_j^{-\epsilon} h$ ,  $f_2 = gh$ . The relationship of being adjacent is to be symmetric, that is, if  $f_1$  and  $f_2$  are adjacent, then so are  $f_2$  and  $f_1$ . We now say that two words  $f$  and  $g$  are *equivalent* and write  $f \sim g$  if there is a finite sequence  $f_1 = f, f_2, \dots, f_m = g$  such that  $f_i$  and  $f_{i+1}$  are adjacent for  $i = 1, 2, \dots, m-1$ . Clearly  $f \sim g$  is a true equivalence relation. All words equivalent to  $f$  form a class which we designate as  $[f]$ .

For any word  $f = a_1 a_2 \cdots a_t$  we define the " $W$ -process," writing,

$$W_0 = 1, \quad \text{the void word}$$

$$W_1 = a_1$$

and for  $i = 1, \dots, t-1$  recursively,

$$\begin{aligned} W_{i+1} &= W_i a_{i+1}, \text{ if } W_i \text{ is not of the reduced form } X a_{i+1}^{-1} \\ &= X, \quad \text{ if } W_i \text{ is of the reduced form } X a_{i+1}^{-1} \end{aligned}$$

We also write  $W(f) = W_t$ . It is not difficult to show that  $W(f)$  is a reduced word, and that if  $f$  is reduced, then  $W(f) = f$ . Furthermore, if  $f \sim g$ , then  $W(f) = W(g)$ . For details see Chapter 7, [16]. The  $W$ -process is mentioned here because it is a simple instance of an effective procedure for deciding an important question: when are two words  $f$  and  $g$  equivalent?

We define a product for classes of words by the rule

$$(2.2) \quad [f][g] = [fg]$$

It is not difficult to show that if  $f_1 \sim f_2$  and  $g_1 \sim g_2$ , then  $f_1 g_1 \sim f_2 g_2$ . Having shown this, it follows that the product rule (2.2) is

well defined. Under this rule the classes of words on generators from the set  $S = \{s_i\}, i \in I$  forms a group, with the void word 1 as the identity. If  $f = a_1 a_2 \cdots a_t$  and  $f^* = a_t^{-1} \cdots a_2^{-1} a_1^{-1}$ , then  $[f][f^*] = [f^*][f] = [1]$  and  $[f^*] = [f]^{-1}$ . The group of these classes is called the *free group* with generators  $S$ , written  $F(S)$ . Note that the property of being free is not given as an abstract property of a group but in terms of a specific set of generators.

Let  $I$  be an index system. We say that elements  $x_i, i \in I$ , of a group  $G$  are *generators* of  $G$  if every element of  $G$  can be written as a finite product of the  $x$ 's and their inverses. Trivially, a group is generated by all its elements. A major theorem is the following.

*Theorem 1. Let  $I$  be an index system and suppose that the group  $G$  is generated by a set of elements  $X: \{x_i\}, i \in I$ . Then if  $F(S)$  is the free group generated by  $S: \{s_i\} i \in I$ , there is a homomorphism  $F \rightarrow G$  determined by  $s_i \rightarrow x_i, i \in I$ .*

PROOF. Let  $f = a_1 \cdots a_t$  be any word on the elements of  $S$ . Consider the element  $g = b_1 \cdots b_t \in G$  where  $b_i = x_j^{\epsilon}$  if  $a_i = s_j^{\epsilon}$ . Then  $f \rightarrow g$  maps every word formed from  $S$  onto an element of  $G$ . Clearly adjacent and therefore equivalent words formed from  $S$  are mapped onto the same element of  $G$ . Hence the mapping  $f \rightarrow g$  is in fact a mapping of elements of the free group  $F(S)$  onto elements of  $G$ . Moreover if  $f_1 \rightarrow g_1, f_2 \rightarrow g_2$ , then  $f_1 f_2 \rightarrow g_1 g_2$ . Hence the mapping  $s_i \rightarrow x_i$  determines a homomorphism of  $F$  onto  $G$ .

The theorems on homomorphisms yield the following as essentially a corollary to the preceding theorem.

*Theorem 2. Every group  $G$  given as generated by a set  $X$  is isomorphic to a factor group of a free group  $F(S)$  generated by a set  $S$  with the same cardinal number of elements as  $X$ .*

Let  $K$  be the kernel of the homomorphism  $F(S) \rightarrow G$  of Theorem 1. Then if  $f = a_1 a_2 \cdots a_t \in K$ , in  $G$  we have  $b_1 b_2 \cdots b_t = 1$ , and conversely. Hence the words  $f(x_i)$  equal to the identity in  $G$  correspond precisely to the words  $f(s_i)$ , which are elements of  $K$  in  $F(S)$ .

*Definition.* A presentation of a group  $G$  is the definition of  $G$  in terms of generators  $X = \{x_i\}, i \in I$ , and relations  $f_j(x) = 1, j \in J$ , where  $I$  and  $J$  are index systems.  $G$  is isomorphic to  $F(S)/K$  where  $F(S)$  is the free group generated by  $S = \{s_i\}, i \in I$ , and  $K$  is the least normal subgroup of  $F(S)$  containing all the words  $f_j(s), j \in J$  under the correspondence  $s_i \rightarrow x_i, i \in I$ .

We say that  $G$  is finitely presented if both the number of generators and the number of relations are finite. Let us note that the isomorphism of  $G$  with  $F(S)/K$  clarifies the question as to which relations are consequences of the given relations  $f_j(x)$ . In  $F(S)$  let us take all the words  $f_j(s)$  and all their conjugates in  $F(S)$ , namely all  $h^{-1}f_j(s)h$  where  $h$  is arbitrary. Then the set of all finite products of elements  $h^{-1}f_j(s)h$  and their inverses is a normal subgroup  $K$  of  $F(S)$ . And clearly any normal subgroup of  $F(S)$  containing the  $f_j(s)$  must contain all of  $h^{-1}f_j(s)h$ , their inverses and finite products of these, and so  $K$ . Thus  $g(x) = 1$  is a consequence of  $f_j(x) = 1$ ,  $j \in J$  in  $G$  if and only if in  $F(S)$   $g(s)$  is a finite product of the  $h^{-1}f_j(s)h$  and their inverses. In  $F(S)$  itself  $K$  is the identity and so  $g(s) = 1$  in  $F(s)$  if and only if the reduced form of  $g(s)$  is the void word.

### 3. THE UNSOLVABLE PROBLEMS

It is not within the scope of this writing to give an adequate account of the proofs that various problems on generators and relations in groups are unsolvable. Nevertheless a listing of some of these problems will be given to show what cannot be done. Thus we can appreciate the value of those methods which solve problems under special conditions which are unsolvable in their unrestricted form.

A rather loose description of unsolvability is the assertion that it is impossible to write a program for a computer which would solve the problem. In this context we think of a computer with an arbitrarily long auxiliary tape which can hold the input data for any individual instance of the problem.

All problems mentioned here deal with countable systems and are questions as to the effective decidability of certain issues. In this sense they are special questions on what the logicians call recursive enumerability. These problems can be cast as questions about the positive integers. A set  $P$  of the positive integers is said to be *recursively enumerable* if there are two effective rules given which list the integers in  $P$  and the integers in the complement of  $P$  respectively. In this definition a more explicit description is needed of what an "effective rule" is, but this requires logical technicalities which do not contribute further to the meaning of this term. We say that the problem as to whether or not an integer is in  $P$  is *decidable* if  $P$  is recursively enumerable. In a practical

sense this may not be satisfactory, for there may be no way of telling how long we must go in the listing of  $P$  and its complement before we are assured of finding a given integer  $n$  in one or the other of these lists, since recursive enumerability gives no indication as to the order of the lists.

The set  $P$  of positive integers  $n$  which can be written in the form  $n = x^3 - y^2$  with  $x, y$  positive integers may not be recursively enumerable, since there is certainly no obvious way of listing them in increasing order. But if a computable method were known for listing the complement, then  $P$  would be recursively enumerable.

The basic problems deal with a finitely presented group  $G$ , where  $G$  is generated by a finite number of generators  $x_1, \dots, x_r$  subject to a finite number of relations  $f_j(x_1, \dots, x_r) = 1, j = 1, \dots, m$ . The *word problem* is the decision as to whether or not a given word  $w(x_1, \dots, x_r)$  is the identity in  $G$ . The unsolvability of the word problem was announced by P. S. Novikov [38, 39] in 1952, and his full proof was given in 1955. Other proofs of the unsolvability of the word problem for groups have been given by J. L. Britton [5] and W. W. Boone [3]. Boone has even shown that there is a group  $G_0$  with two generators and thirty-two defining relations for which the word problem is unsolvable. The thirty-two relations could presumably be written out, although they would be enormously long.

As a consequence of the unsolvability of the word problem, a number of other problems can be shown unsolvable. G. Baumslag, W. W. Boone, and B. H. Neumann [2] have obtained the following results:

*Theorem 3. There is a finitely presented group  $G_0$  such that no effective procedure exists to determine whether a word given in generators of  $G_0$  represents:*

1. *an element of the center of  $G_0$ ;*
2. *an element permutable with a given element of  $G_0$ ;*
3. *an  $n$ th power, where  $n > 1$  is an integer;*
4. *an element whose class of conjugates is finite;*
5. *an element of a given subgroup of  $G_0$ ;*
6. *a commutator;*
7. *an element of finite order  $> 1$ .*

They show that there are many unsolvable problems on finitely presented groups, and follow Rabin [41] to prove a very general theorem.

*Theorem 4.* Let  $P$  be an algebraic property of groups (i.e., one that is shared by all isomorphic copies of any group that has it), and assume (1) that there is a finitely presented group that has  $P$ , and (2) that there is an integer  $n$  such that no free group  $F_r$  of rank  $r \geq n$  has  $P$ ; then there is a finitely presented group  $G_P$  such that no effective procedure exists to determine whether the elements represented by a finite set of words in given generators of  $G_P$  generate a subgroup with  $P$ .

Graham Higman [24] has made a further study of finitely generated groups and has shown the following:

*Theorem 5.* A finitely generated group can be embedded in a finitely presented group if and only if it is recursively presented.

This general result is proved independently of the earlier theorems and yields incidentally a new proof of the unsolvability of the word problem.

From Theorem 4 the authors derive the following:

*Theorem 6.* There are finitely presented groups  $G$  (depending on the property considered) such that no effective procedure exists to determine whether the elements represented by a finite set of words generate a subgroup of  $G$  that is (1) trivial, (2) finite, (3) free, (4) locally infinite, (5) cyclic, (6) Abelian, (7) nilpotent, (8) solvable, (9) simple, (10) directly indecomposable, (11) freely decomposable, (12) a group with solvable word problem; and so on *ad nauseam* (sic!).

These theorems show that along with the word problem, a host of other problems are unsolvable for finitely presented groups. Since the word problem is solvable for finitely presented Abelian groups, we can decide if an element is in the derived group of a finitely presented group.

This state of affairs raises a major question.

*Problem 1.* Are there broad classes of groups for which the word problem is solvable? Can such groups be easily recognized?

There are certainly classes of groups for which the word problem is solvable. Magnus [32] has solved the word problem for any group with a single defining relation. Tartakovskii [45, 46] has given a solution for the word problem when the defining relations do not combine readily two at a time.

## 4. FREE GROUPS, FREE PRODUCTS,

In Section 2 we gave a definition of free groups. If  $G_i$  is a set of groups indexed by letters  $i \in I$ , where  $I$  is a well-ordered index system, we may define the *free product*  $\prod_i G_i$  of the groups  $G_i$  in a manner similar to that used in defining free groups. Consider words (or strings)

$$(4.1) \quad a_1 a_2, \dots, a_t$$

which are either void (written 1) or in which each  $a_i$ ,  $i = 1, \dots, t$  is an element of some  $G_j$ . For these words we define two kinds of *elementary equivalence* (or adjacency):

$$E1: \quad a_1 a_2 \dots a_{i-1} a_i a_{i+1} \dots a_t \sim a_1 \dots a_{i-1} a_{i+1} \dots a_t$$

if  $a_i$  is the identity element of some  $G_j$ .

$$E2: \quad a_1 a_2 \dots a_{i-1} a_i a_{i+1} a_{i+2} \dots a_t \\ \sim a_1 \dots a_{i-1} a_i^* a_{i+2} \dots a_t$$

if  $a_i$  and  $a_{i+1}$  are in the same  $G_j$  and  $a_i a_{i+1} = a_i^*$  in  $G_j$ . These elementary equivalences are taken to be symmetric relations. Two words  $x$  and  $y$  are said to be *equivalent* if there is a finite sequence  $x_1 = x, x_2, \dots, x_n = y$  with  $x_i$  and  $x_{i+1}$  elementary equivalent for  $i = 1, \dots, n-1$ . We say that a word  $w = a_1 a_2 \dots a_t$  is *reduced* if it is either void or if (1) no  $a_i$ ,  $i = 1, \dots, t$  is the identity of its group  $G_j$ , and (2)  $a_i, a_{i+1}$ ,  $i = 1, \dots, t-1$ , belong to different groups  $G_j$ . We call  $t$  the *length* of  $w$ . As in Section 2 we can define a  $W$ -process for words and easily show that in every class of equivalent words there is exactly one reduced word. If we let  $[f]$  denote the class of words equivalent to a particular  $f$ , then the rule  $[f][g] = [fg]$  gives a multiplication for the classes which can be shown to be well defined. Under this multiplication the classes of equivalent words form a group in which the void word is the identity. This group  $G$  is the *free product* of the groups  $G_i$ ,  $i \in I$  and we write

$$(4.2) \quad G = \prod_i^* G_i, \quad i \in I$$

Free products derive their chief significance from the following theorem whose proof is essentially the same as that of Theorem 1.

*Theorem 7. Let  $G$  be a group generated by the elements of a family of subgroups  $H_i$ ,  $i \in I$  where the  $H_i$  are isomorphic to groups  $G_i$ .*

*Then  $G$  is a homomorphic image of the free product  $Q = \prod_i^* G_i$ ,  $i \in I$ .*

We note in passing that a free group is the special case of free product in which each  $G_i$  is an infinite cyclic group. But there are certain differences in their treatment. Thus if  $F_2$  is the free group generated by elements  $x$  and  $y$  the element  $x^{-2}y^{-1}x^2y$  has its length  $t = 6$ , but in the free product of  $\{x\}$  and  $\{y\}$  its length is  $t = 4$ . We also note that the existence of free products shows that in a general group  $G$  relations  $f_j(x) = 1$  on a subset  $x_1, \dots, x_r$  of its generators and other relations  $f_k(y) = 1$  on a disjoint subset of generators  $y_1, \dots, y_s$  are quite independent of each other in a sense which may be made precise by reference to the free product.

Suppose we have a family of groups  $G_i$ ,  $i \in I$ , and that various subgroups of different  $G$ 's are isomorphic to each other. What group  $G$  is most freely generated by the  $G_i$  with these subgroups identified? This is the *amalgamated product* developed initially by Hannah Neumann [35]. We wish to specify in advance the intersection of  $G_i$  and  $G_j$  for  $i, j \in I$ . This will be a subgroup of  $G_i$  and an isomorphic subgroup of  $G_j$ . Hence we must specify not only the two subgroups but isomorphism between them. We are given subgroups  $U_{ij}$  and isomorphisms  $I_{ij}$

$$(4.3) \quad \begin{aligned} U_{ij} &\subseteq G_i, \quad i, j \in I \\ U_{ji} &\subseteq G_j \\ I_{ij}(U_{ij}) &= U_{ji} \end{aligned}$$

But from the property of intersections we require

$$(4.4) \quad \begin{aligned} G_i \cap G_j \cap G_k &= (G_i \cap G_j) \cap (G_i \cap G_k) \\ &= (G_j \cap G_i) \cap (G_j \cap G_k) \\ &= (G_k \cap G_i) \cap (G_k \cap G_j) \end{aligned}$$

For this to be satisfied, the subgroups and isomorphisms of (4.3) must satisfy

$$(4.5) \quad I_{jk}[I_{ij}(U_{ij} \cap U_{ik})] = I_{ik}(U_{ij} \cap U_{ik}), \quad i, j, k \in I$$

The relations (4.5) are sufficient to insure that intersections of more than three  $G$ 's will be consistent with the identifications.

If  $Q = \prod_i G_i$  is the free product of the  $G$ 's, let  $K$  be the normal subgroup of  $Q$  generated by all elements  $u_{ij}u_{ji}^{-1}$  and their conjugates where  $I_{ij}(u_{ij}) = u_{ji}$ ,  $i, j \in I$ . Then  $Q/K$  is the desired group most freely generated by the  $G_i$  with the identifications desired. But we have no guarantee that these identifications do not have mysterious consequences. For example we might find that  $Q/K$  is trivial. Thus if we take the three groups

$$(4.6) \quad \begin{aligned} G_1 &= \{x_1, y_1\}, & x_1^{-1}y_1x_1 &= y_1^2 \\ G_2 &= \{y_2, z_2\}, & y_2^{-1}z_2y_2 &= z_2^2 \\ G_3 &= \{z_3, x_3\}, & z_3^{-1}x_3z_3 &= x_3^2 \end{aligned}$$

it can be shown that each of the elements  $x_1, y_1, y_2, z_2, z_3, x_3$  generates an infinite cyclic group. Hence the identifications  $x_1 = x_3$ ,  $y_1 = y_2$ ,  $z_2 = z_3$  involve the amalgamation of isomorphic subgroups. But with these amalgamations we have the group of (2.1)

$$(4.7) \quad G = \{x, y, z\}, \quad x^{-1}yx = y^2, \quad y^{-1}zy = z^2, \quad z^{-1}xz = x^2$$

which is trivial. Let us follow B. H. Neumann [34] to prove this. From the first relation

$$(4.8) \quad y^{-1}xy = xy^{-1}$$

and so

$$(4.9) \quad y^{-i}xy^i = xy^{-i}$$

Similarly, the second relation gives

$$(4.10) \quad z^{-i}yz^i = yz^{-i}$$

Transforming the third relation by  $y$ , we have

$$(4.11) \quad y^{-1}z^{-1}xzy = y^{-1}x^2y$$

This is

$$(4.12) \quad y^{-1}z^{-1}y \cdot y^{-1}xy \cdot y^{-1}zy = y^{-1}x^2y$$

or, using (4.8)

$$(4.13) \quad \begin{aligned} z^{-2}xy^{-1}z^2 &= (xy^{-1})^2 = x(y^{-1}xy)y^{-2} \\ &= x(xy^{-1})y^{-2} = x^2y^{-3} \end{aligned}$$

Now using (4.10) and the third relation to alter the left hand side of (4.13), we have

$$(4.14) \quad x^4 z^2 y^{-1} = x^2 y^{-3}$$

From this

$$(4.15) \quad z^2 = x^{-2} y^{-2}$$

Also  $z^2 = y^{-1} z y$  gives

$$(4.16) \quad z y z^{-1} = y z$$

and so

$$(4.17) \quad z^2 y z^{-2} = z y z^{-1} \cdot z = y z^2$$

Substituting (4.15) into (4.17) gives

$$(4.18) \quad (x^{-2} y^{-2}) y^2 (y^2 x^2) = y z^2$$

or

$$(4.19) \quad x^{-2} y^2 x^2 = y z^2$$

and so

$$(4.20) \quad y^4 = y z^2, \quad y^3 = z^2$$

Hence  $y$  and  $z^2$  commute and the second relation gives

$$(4.21) \quad z = z^2$$

From this it is immediate that  $z = 1$  and in turn  $x = 1, y = 1$ , and  $G$  is trivial.

We shall say that the amalgamated product of the  $G_i, i \in I$  with identifications  $U_{ij} = U_{ji}$  exists, provided that there are no unwanted

identifications. Explicitly we require, (1) that in  $Q = \prod_i G_i$  with

the  $K$  the kernel,  $K = \{w^{-1}(u_{ij}u_{ji}^{-1})w\}, G_i \cap K = 1, \text{ all } i \in I;$  and (2) if in  $Q g_i \in G_i, g_j \in G_j$  and  $g_i g_j \in K$ , then for an appropriate  $u_{ij} \in U_{ij}$  and  $u_{ji} = I_{ij}(u_{ij})$  we have  $g_i = u_{ij}, g_j = u_{ji}^{-1}$ . Under these circumstances we say the amalgamated product  $G$  exists, and we write

$$(4.22) \quad G = \prod_i^* G_i : U_{ij} = U_{ji}, i, j \in I$$

In one simple case the conditions (1) and (2) are always satisfied, and the amalgamated product exists. This is the case in which all the  $U_{ij}$  are isomorphic to the same group  $U$ . This case was treated by Schreier [44] in 1927. Let  $x_{i1} = 1, \dots, x_{is}, \dots$ , be a fixed set of representatives of the left cosets  $Ux = Ux_{is}$  in  $G_i$ . Then any element  $g$  of  $G = \prod_i G_i: U_{ij} = U$  can be put into a canonical form

$$(4.23) \quad g = ua_1a_2 \cdots a_t$$

with  $u \in U$ , each  $a_j = x_{is} \neq 1$  a coset representative and  $a_j$  and  $a_{j+1}j = 1 \cdots t - 1$  from different  $G$ 's. Furthermore, these are "reduced" words in the sense that two words of this form are the same group element only if they are identical. Note that this guarantees the existence of an amalgamated product of two groups.

For each  $i \in I$  let  $U_i = \{U_{ij} | j \neq i\}$ . It is clear then that if the amalgamated product of the  $G_i$  exists, the amalgamated product of the  $U_i$  exists. Hannah Neumann [35] has shown conversely that if the amalgamated product of the  $U_i$  exists, also that of the  $G_i$  exists.

*Problem 2. Find conditions for the existence of an amalgamated product of three groups.*

*Problem 3. Find a canonical form for the elements in an amalgamated product when it exists.*

Let  $G$  be a group generated by elements  $s_1, s_2, \dots, s_r$  (or  $s_i$ ,  $i \in I$  if the number is infinite). What can we say about generators of a subgroup  $H$ ? Let  $x_1 = 1, x_2, \dots, x_n$  (allowing  $n$  to be infinite if necessary) be representatives of the left cosets of  $H$  in  $G$  thus

$$(4.24) \quad G = H + Hx_2 + \cdots + Hx_n$$

Furthermore let us define a function  $\Phi(g)$  for  $g \in G$  by the rule

$$(4.25) \quad \Phi(g) = x_i \text{ if } g \in Hx_i$$

We note that  $\Phi$  is a decision function since  $\Phi(g) = 1$  if  $g \in H$ . Thus by the unsolvability of the word problem for finitely presented groups there is a normal subgroup  $H$  of a finitely generated free group  $F$  for which the function  $\Phi$  is not computable. Hence in using (4.24) and (4.25) we are already passing over some funda-

mental difficulties. Let  $\Phi$  be the function  $\phi$  restricted to arguments  $g = s_i x_j^\epsilon, \epsilon = \pm 1$ .

$$(4.26) \quad \phi(s_i s_j^\epsilon) = \Phi(s_i s_j^\epsilon)$$

**Theorem 8.** *If  $s_i, i \in I$  are the generators of  $G$  and  $x_j, j \in J$  are representatives of left cosets of a subgroup  $H$  of  $G$  then  $u_{ij} = x_i s_j \phi(x_i s_j)^{-1}$  are generators of  $H$ .*

**PROOF.** Let us first observe that the function  $\phi$  satisfies three simple properties.

$$(4.27) \quad \begin{aligned} (1) & \quad \phi(x_i s_j^\epsilon) \text{ is an } x_k \\ (2) & \quad \text{If } x_i s_j^\epsilon \text{ is an } x_k, \text{ then } \phi(x_i s_j^\epsilon) = x_i s_j^\epsilon \\ (3) & \quad \phi[\phi(x_i s_j^\epsilon) s_j^{-\epsilon}] = x_i \end{aligned}$$

The first two properties are immediate. For the third we merely note that if  $\phi(x_i s_j^\epsilon) = x_k$ , then  $x_i s_j^\epsilon = h x_k$  for some  $h \in H$  whence  $x_k s_j^{-\epsilon} = (h^{-1} x_i s_j^\epsilon) s_j^{-\epsilon} = h^{-1} x_i$ , and so  $\phi(x_k s_j^{-\epsilon}) = x_i$ . Now suppose that

$$(4.28) \quad a_1 a_2 \cdots a_t = h \in H, \quad \text{each } a_m \text{ some } s_{j'}^\epsilon$$

Then identically

$$(4.29) \quad h = [\Phi(1) a_1 \Phi(a_1)^{-1}] [\Phi(a_1) a_2 \Phi(a_1 a_2)^{-1}] \cdots \\ [\Phi(a_1 \cdots a_{m-1}) a_m \Phi(a_1 \cdots a_m)^{-1}] \cdots \\ [\Phi(a_1 \cdots a_{t-1}) a_t \Phi(a_1 \cdots a_t)^{-1}]$$

where we note that  $\Phi(1) = 1$  and also  $\Phi(a_1 \cdots a_t) = \Phi(h) = 1$ . Thus  $h$  is a product of factors of the form

$$v = \Phi(a_1 \cdots a_{m-1}) a_m \Phi(a_1 \cdots a_m)^{-1}$$

Let  $a_1 \cdots a_{m-1} = h_1 x_i$  for  $h_1 \in H$  and  $x_i$  a coset representative and let  $a_m = s_{j'}, s_{j'}$  a generator. Then if  $x_i s_{j'}^\epsilon = h_2 x_k, h_2 \in H, x_k$  a coset representative we have  $\Phi(a_1 \cdots a_{m-1}) = x_i \Phi(a_1 \cdots a_m) = x_k = \Phi(x_i s_{j'}^\epsilon)$ . Also  $v = h_1 x_i s_{j'}^\epsilon s_k^{-1} = h_1 h_2 \in H$ . Thus  $H$  is generated by elements of the form  $v = x_i s_{j'}^\epsilon \phi(x_i s_{j'}^\epsilon) = x_i s_{j'}^\epsilon x_k^{-1}$ . But then  $v^{-1} = x_k s_{j'}^{-\epsilon} x_i^{-1} = x_k s_{j'}^{-\epsilon} \phi(x_k s_{j'}^{-\epsilon})$  using property (3) of the  $\phi$  function. Hence each factor in (4.29) or its inverse is of the form  $u_{ij} = x_i s_j \phi(x_i s_j)^{-1}$ , and so these elements generate  $H$ .

There are two immediate consequences of this theorem and its proof: (1) if  $G$  is finitely generated and  $H$  is of finite index in  $G$ , then  $H$  is finitely generated and (2)  $\Phi$  is computable from  $\phi$ .

At this stage there has been no restriction on the choice of coset representatives except for choosing the identity as the representative for  $H$ . We can expect to gain certain advantages by choosing the representatives systematically. Let us follow Schreier [44] and ask that the representatives be what we call, after him, a Schreier System  $S$ .

*Definition.* A set of words on generators  $s_1, \dots, s_r$  is a Schreier System  $S$ ,

1. if for  $a_1 \cdots a_t \in S$ ,  $a_1 \cdots a_t$  is reduced, that is,  $a_i \neq 1$ ,  $i = 1, \dots, t$ ,  $a_i a_{i+1} \neq 1$ ,  $i = 1, \dots, t-1$ ;
2. if  $a_1 a_2 \cdots a_t \in S$ , then  $a_1 \cdots a_{t-1} \in S$ .
3. if  $a_1 a_2 \cdots a_t \in S$ , then  $a_2 \cdots a_t \in S$ .

If 1 and 2 hold we call  $S$  a Schreier System. If 1, 2, and 3 hold we call  $S$  a two-sided Schreier system.

*Theorem 9.* If  $H$  is a subgroup of  $G$ , where  $G$  is generated by  $s_i$ ,  $i \in I$  we may choose the representatives of the left cosets of  $H$  in  $G$  to be a Schreier system, and if  $H$  is normal in  $G$  to be a two-sided Schreier system.

The proof of this is easy and will not be given in detail. If the  $s_i$ ,  $i \in I$  and their inverses are well ordered, we may determine a well ordering of words  $a_1 \cdots a_t$  in the  $s_i$  and their inverses by ordering words first in terms of their reduced length  $t$  and for  $a_1 \cdots a_t$  and  $b_1 \cdots b_t$  both reduced of length  $t$  ordering on the first  $a_i \neq b_i$ ,  $i = 1, \dots, t$ . In terms of this alphabetical ordering, if we choose as the representative of a coset the element which as a word is earliest in the ordering, the theorem is easily proved.

Let us suppose that the representatives  $x_i$  of the left cosets of  $H$  in  $G$  are a Schreier system. Then if  $g = a_1 a_2 \cdots a_t$  is any word on the generators of  $G$ , let us define its  $\Phi$  form  $k(g)$  by

$$(4.30) \quad k(g) = (\Phi(1)a_1\Phi(a_1)^{-1}) \cdots \\ [\Phi(a_1 \cdots a_{m-1})a_m\Phi(a_1 \cdots a_m)^{-1}] \\ (\Phi(a_1 \cdots a_{t-1})a_t\Phi(a_1 \cdots a_t)^{-1}) \cdot \Phi(a_1 \cdots a_t)$$

Here trivially  $k(g) = g$  as an element of  $G$ . Also from property (3) of the  $\phi$  function,  $k(g_1 s_j s_j^{-1} g_2) \sim k(g_1 g_2)$ , being adjacent words in the  $u$ 's, and so the reduced  $\Phi$  form of  $g$  depends only on the reduced form of  $g$ . We easily prove the following lemma:

*Lemma 1.* If the representatives  $y_i$  form a Schreier system, then either a  $v = x_i s_j^e \phi(x_i s_j^e)^{-1}$  is in reduced form or its reduced form is the identity.

Let us delete from the  $u_{ij}$ 's those which are trivial, that is, those whose reduced forms are the identity, and call  $U$  the remaining set of  $u_{ij}$ 's, and  $U^{-1}$  the set of inverses of these  $u_{ij}$ 's.

*Lemma 2.* Let  $g = A_1 A_2 \cdots A_t$  be a word, possibly void, where each  $A_i$  is in  $U$  or  $U^{-1}$ . If each  $A$  is written in its reduced form in terms of the  $s_i$ 's, then omitting trivial  $u$ 's the  $\Phi$  form of  $g$  is  $A_1 A_2 \cdots A_t$ .

The proof of this is straightforward and will be omitted. It depends on the observation that the  $\Phi$  form of an  $A$  belonging to  $U$  or  $U^{-1}$  involves trivial  $u$ 's and itself.

*Lemma 3.* If  $A_1 A_2 \cdots A_t = 1$  where each  $A_i$  belongs to  $U$  or  $U^{-1}$  then  $A_1 A_2 \cdots A_t$  is the  $\Phi$  form of a word  $a_1 a_2 \cdots a_w$ , each  $a_i = s_j^e$  which is the identity, omitting trivial  $u_{ij}$ 's in the  $\Phi$  form.

This lemma is an immediate consequence of the preceding. Of course when a word in the  $u_{ij}$  and their inverses is equal to 1, replacement of the  $u$ 's by their given forms in the  $s_j^e$  yields a word equal to 1. The force of this lemma is the converse. We can now state our main theorem.

*Theorem 10.* Let a group  $G$  have generators  $s_a$ ,  $a \in I$  and defining relations  $f_j(s) = 1$ ,  $j \in J$ . Let  $H$  be a subgroup of  $G$  and let  $x_i$  be a Schreier system of representatives of left cosets of  $H$  in  $G$ . Write  $u_{ij} = x_i s_j \phi(x_i s_j)^{-1}$ . The nontrivial  $u$ 's are generators of  $H$ . Defining relations for  $H$  are the  $\Phi$  forms of  $k(x_i f_j(s) x_i^{-1}) = 1$ , omitting trivial  $u$ 's.

**PROOF.** Everything has already been proved except the statement about the defining relations for  $H$ . From Lemma 3 a word equal to 1 in the  $u$ 's and their inverses is the  $\Phi$  form (omitting trivial  $u$ 's) of a word in the  $s_j^e$  equal to 1. Such a word is a product of word  $g^{-1} f_j(s) g$  and their inverses, but each of these is separately the identity. Hence the  $\Phi$  forms of  $g^{-1} f_j(s) g$  are defining relations for  $H$ . There is a redundancy here, however. We can write  $g^{-1} = h x_i$  for some  $h \in H$  and a coset representative  $x_i$ . Then  $k[g^{-1} f_j(s) g] = k[h x_i f_j(s) x_i^{-1} h^{-1}] = k(h) k[x_i f_j(s) x_i^{-1}] k(h)^{-1}$  where

$k(h^{-1}) = k(h)^{-1} \in Hh$  being an element of  $H$ . Thus the  $\Phi$  forms  $k[x_i f_j(s)x_i^{-1}] = 1$  are defining relations for  $H$ .

Thus several major theorems are now corollaries of what has been proved.

*Theorem 11. Every subgroup of a free group is itself a free group.*

This theorem follows from the preceding theorem since the defining relations of the subgroup all reduce to the identity.

*Theorem 12. A subgroup of finite index in a finitely presented group is itself finitely presented.*

PROOF. Here the number of  $x_i$ 's,  $s_j$ 's, and  $f_a(s) = 1$  is finite, and so the number of  $u_{ij}$ 's and  $k[x_i f_a(s)x_i^{-1}] = 1$  is also finite.

*Theorem 13. If  $G$  is generated by  $s_1, \dots, s_r$  and if  $H$  is of index  $n$ , the number of nontrivial  $u_{ij}$  is  $1 + (r - 1)n$ . The total length of the non-trivial  $u_{ij}$  is  $(2L + n)r - 2L$  where  $L$  is the total length of the  $x$ 's.*

PROOF. There are  $rn$   $u_{ij}$ 's including the trivial ones. It is merely a matter of deciding how many trivial ones there are. From Lemma 1,  $x_i s_j \phi(x_i s_j)^{-1} = 1$  if either  $x_i s_j$  is in reduced form and is an  $x_k$  or if the reduced form of  $x_i$  ends in  $s_j^{-1}$ . Hence for fixed  $j$  the number of trivial  $x_i s_j \phi(x_i s_j)^{-1}$  is the number of coset representatives ending in  $s_j$  or  $s_j^{-1}$ . If we take all  $j = 1, \dots, r$  into account, this accounts for every coset representative exactly once except for the identity. Hence there are  $n - 1$  trivial  $u_{ij}$ 's, and the number of nontrivial  $u_{ij}$ 's is  $rn - (n - 1) = 1 + (r - 1)n$ . Similarly, if  $L$  is the total length of the  $x_i$ , the total length of all  $u_{ij}$  before reduction is  $(2L + n)r$  and the total length of the nontrivial  $u_{ij}$  is  $(2L + n)r - 2L$ .

The foregoing theorems stem from papers by Schreier [44] Hall and Rado [17] and others [15].

Let  $F$  be a free group with  $r$  generators  $x_1, x_2, \dots, x_r$ . We say that  $F$  is of rank  $r$ . It is not difficult to verify that  $y_1 = x_1 x_2$ ,  $y_2 = x_2, \dots, y_r = x_r$  are also free generators for  $F$ . Thus the free generators are not uniquely determined. But we can show that the number of free generators for  $F$  is an invariant. Let  $T$  be the subgroup of  $F$  generated by all squares of elements. Then as  $x^{-1}y^{-1}xy = (x^{-1}y^{-1}x)^2x^{-2}(xy)^2$ , we see that  $T$  contains all commutators. Here  $T$  is a normal subgroup of  $F$  (in fact fully invariant), and  $F/T$  is the elementary Abelian group of order  $2^r$  if  $r$  is

finite, and is infinite if  $r$  is infinite. If  $F$  has infinitely many generators, the cardinality of  $F$  is the same as that of its number of generators. Hence the number of generators  $1 + (r - 1)n$  given in Theorem 13 is the best possible value if  $G$  is free. But if  $G$  is not free, it often happens that fewer generators will suffice for  $H$  and there may be a redundancy in the defining relations for  $H$  as given by Theorem 10. This situation will certainly arise if one of the nontrivial  $u_{ij}$ 's is the identity.

*Problem 4.* Let  $G$  be a finitely presented group, and  $H$  a subgroup of finite index. Find, under suitable restrictions if necessary, smaller values for the number of generators and defining relations for  $H$ .

As an example suppose that  $G$  has generators  $x$  and  $y$  satisfying  $x^4 = y^4 = 1$ , and that  $H = G'$ . Here  $[G:H] = 16$ . Our theorems tell us that  $H$  has seventeen generators and thirty-two relations. But in fact these numbers may be reduced to express  $H$  as a group with nine generators and no relations, that is  $H$  is a free group of rank 9. Group  $G$  is a free product of two cyclic groups of order 4. A paper by MacLane [31] gives some information on Problem 4 in this case.

The preceding method of considering subgroups, which we shall call the Schreier method, depends on consideration of coset representatives. It has already been observed that this method assumes a knowledge of which elements are in which coset, although we know that in general this is undecidable for finitely presented groups.

A second approach to the generation of subgroups deals more directly with the elements of the subgroup. This we shall call the Nielsen method, since Nielsen [36] in 1921 first used this procedure. We shall give here a refinement of the writer [15].

Let  $S = \{s_i\}$ ,  $i \in I$  be the generators of a group  $G$  where we assume  $I$  to be well ordered. Then we may well order the generators and their inverses putting  $s_i < s_i^{-1} < s_j < s_j^{-1}$  if  $i < j$  in  $I$ . We introduce an *alphabetical ordering*  $g_1 < g_2$  on elements of  $G$  by the following rules. Write  $g_1 = a_1 a_2 \cdots a_t$ , each  $a_i = s_j^{\epsilon}$ ,  $\epsilon = \pm 1$ ,  $g_2 = b_1 b_2 \cdots b_u$ . Then, (1)  $g_1 < g_2$  if the lengths are unequal,  $L_S(g_1) = t < u = L_S(g_2)$  (2) if  $L_S(g_1) = L_S(g_2) = t$ ,  $g_1 = a_1 a_2 \cdots a_t$ ,  $g_2 = b_1 b_2 \cdots b_t$  we put  $g_1 < g_2$  if  $a_1 = b_1$ ,  $a_2 = b_2$ ,  $\dots$ ,  $a_{i-1} = b_{i-1}$ , but  $a_i < b_i$ . As it stands this is strictly an ordering on words in the generators of  $G$ . It is a well ordering, and so we may take as a canonical form for an element  $g$  the earliest

word in the ordering to which it is equal. If  $G$  is a free group, this form is merely the reduced form of  $g$ . If  $G$  is a free product of groups  $G_i$ , we take as generators of  $G$  all elements of the free factors  $G_i$  except the identity. Here again the earliest word representing an element  $g$  is its reduced form. In general the determination of the canonical form of an element presents difficulties. Indeed the unsolvable problems tell us that this determination is sometimes impossible.

In terms of the alphabetical ordering of words in the generators  $S = \{s_i\}$ ,  $i \in I$  and their inverses we may define a *semialphabetical* ordering. For a word  $g$  of even length we write  $g$  in the form  $g = \alpha\beta^{-1}$  with  $L_S(\alpha) = L_S(\beta)$ . For a word  $g$  of odd length we write  $g$  in the form  $g = \alpha a \beta^{-1}$  with  $L_S(\alpha) = L_S(\beta)$  and  $a = s_j^{\epsilon}$ . Here we order the words by the rules, saying  $g_1$  precedes  $g_2$  if (1)  $L_S(g_1) < L_S(g_2)$ . For words of equal length we order successively on (2) the alphabetical order of  $\alpha$ , (3) the alphabetical order of  $\beta$ , and (4) the order of the central  $a$ .

In a free group or a free product the reduced form of an element will be the earliest word equal to it and so its canonical form.

The *Nielsen process* for a finite set of elements in a group  $G$ :  $g_1, \dots, g_m$ , involves the following reductions:

1. Delete a  $g_i = 1$ .
2. If  $g_i^{-1}$  precedes  $g_i$ , replace  $g_i$  by  $g_i^{-1}$ .
3. If one of  $g_i^{\epsilon} g_j^{\eta}$ ,  $g_j^{\eta} g_i^{\epsilon}$ ,  $\epsilon = \pm 1$ ,  $\eta = \pm 1$ ,  $j \neq i$ , precedes  $g_i$ , replace  $g_i$  by this word.

Here "precedes" refers to the semialphabetical ordering. Since the words are well ordered there are only a finite number of successive reductions possible. Furthermore, at every stage the elements retained generate the same group. Nielsen's original paper showed that elements in a free group which could not be reduced were themselves generators of a free group. This showed that every finitely generated subgroup of a free group was itself free. The restriction of this proof to finitely generated subgroups is easily removed. Suppose  $H$  is a subgroup of  $G$ . Our ordering is a well ordering of the elements of  $G$ . Let us choose a set  $K$  of elements from  $H$  by the rule that  $K$  contains every element  $h \neq 1$  of  $H$  which is not in the subgroup of  $H$  generated by elements of  $H$  preceding  $h$  in the ordering. It is easy to show that  $K$  generates  $H$  and that every finite subset of  $K$  is *Nielsen reduced*, that is, that no one of the three reductions is possible. Hence if  $G$  is free, the original Nielsen

proof may be used to show that  $H = F(K)$ , the free group generated by the elements of  $K$ .

Now suppose that  $G = \prod_i^* A_i$ ,  $i \in I$  is a free product and that we consider the elements of the  $A_i$  (excluding the identity) as generators of  $G$ . As just shown, let  $H$  be a subgroup of  $G$  and then  $h = \{K\}$ , where  $K$  is defined by

$$(4.31) \quad K = [h \in H | h \neq 1, h \notin \{h'\} | h' \in H, h' < h]$$

where  $h' < h$  means that  $h'$  precedes  $h$  in the semialphabetical ordering. Furthermore, every finite subset of  $K$  is Nielsen reduced. For  $k \in K$ ,  $k$  is one of the forms

$$(4.32) \quad \begin{aligned} k &= \alpha\beta^{-1} \\ k &= \alpha a\beta^{-1}, \beta \neq \alpha, a \in A_j \\ k &= \alpha a\alpha^{-1}, a \in A_j \end{aligned}$$

The third form we call a transform. From the definition of  $K$  it is possible to draw certain conclusions as to the way in which products of  $k$ 's combine. If  $g = a_1 a_2 \cdots a_i a_{i+1} \cdots a_t$  and  $a_i, a_{i+1}$  belong to the same free factor  $A_j$ , we say that  $a_i$  and  $a_{i+1}$  cancel if  $a_i a_{i+1} = 1$  and amalgamate if  $a_i a_{i+1} = a_i^* \neq 1$ .

*Lemma 4.* In a product  $k_1^* k_2^*$  or  $k_2^* k_1^*$ ,  $k_1, k_2 \in K$ ,  $\eta = \pm 1$ , then, (1) if  $k_1 = \alpha\beta^{-1}$ ,  $\beta$  does not cancel and if  $\alpha$  cancels, then the adjacent term of  $\beta^{-1}$  does not amalgamate; (2) if  $k_1 = \alpha a\beta^{-1}$ ,  $\alpha \neq \beta$ ,  $\alpha$  and  $a$  do not cancel and if  $\beta$  cancels then  $a$  does not amalgamate; (3) if  $k_1 = \alpha a\alpha^{-1}$ ,  $a \in A_j$  and if  $k_2^* = \alpha a_1 \sigma$  with,  $a, a_1 \in A_j$ , then  $a_1$  is the earliest element in the coset  $B_j a_1$  where  $B_j = \{a_i\}$ ,  $\alpha a_i \alpha^{-1} \in K$ ,  $a_i \in A_j$ .

The proof of this lemma is straightforward and may be found in the writer's paper [13]. It is sufficient to show that  $k$ 's of the first two forms in (4.32) are free generators of a free group and the generators of the third form generate conjugates of subgroups of the  $A_j$ . Thus we are able to describe exactly subgroups of free products. The original proof by Kurosch used a complicated double induction.

*Theorem 14.* (Kurosch). A subgroup  $H \neq 1$  of a free product

$$G = \prod_i^* A_i, i \in I$$

is itself a free product

$$H = F * \prod_j^* x_j^{-1} U_j x_j$$

where  $F$  is a free group and each  $x_j^{-1} U_j x_j$  is the conjugate of a subgroup  $U_j$  of one of the free factors  $A_i$  of  $G$ .

MacLane [31] has described the free factors of Theorem 14 more precisely. For each  $i \in I$  the  $x_j$ 's such that  $H \cap x_j^{-1} A_i x_j \neq 1$  are a subset of appropriately chosen representatives of  $H$ - $A_i$  double cosets. If  $[G:H]$  is finite the number of free generators of  $F$  is  $1 - [G:H] + \sum_i ([G:H] - m_i)$  where  $m_i$  is the number of  $H$ - $A_i$  double cosets in  $G$ .

A third method is a topological one. If  $G$  is the fundamental group of a complex, a subgroup  $H$  of  $G$  is the fundamental group of a covering of  $R$ . Baer and Levi [1] have used this approach to prove the Kurosch Theorem 14 on subgroups of free products.

Of free groups, free products, and amalgamated products, the free groups are the simplest and most fundamental. We shall give a summary of further results on free groups, some of which may be generalized to free or amalgamated products.

Not every Schreier system may be the coset representatives of a subgroup of a free group. For example, let  $F_2$  be the free group generated by  $x$  and  $y$ , and  $V$  the system  $1, x, x^2, \dots, x_j^i, \dots$ . If these are the representatives of cosets of a subgroup  $H$ , the element  $x^{-1}$  cannot belong to  $Hx^j$ , for then  $x^{j+1}$  is an element of  $H$  whereas we have assumed that  $Hx^{j+1} \neq H$ . The condition that a Schreier system  $V$  be a set of representatives for some subgroup of  $F$  is the following [see Schreier condition]: Let  $V$  be a Schreier system in a free group  $F$  with generators  $s_i, i \in I$ . Let  $M(s_i)$  and  $M(s_i^{-1})$  be the cardinality of the sets of elements of  $V$  not of the form  $xs_i, x \in V$  or  $xs_i^{-1}, x \in V$  respectively. Then  $M(s_i) = M(s_i^{-1})$ , every  $i \in I$  is a necessary and sufficient condition that  $V$  be the coset representatives of some subgroup  $H$  of  $F$ . This comes rather naturally from the fact that the mappings  $x_j \rightarrow \phi(xs_i), x_j \rightarrow \phi(xs_i^{-1})$  where the  $x_j$ 's are coset representatives must be one-to-one on  $V$ . It is easy to show that in a Schreier system the cardinality of the numbers  $N(s_i)$  and  $N(s_i^{-1})$  of elements of  $V$  which are of the forms  $x_j s_i$  and  $x_j s_i^{-1}$  are the same. Thus the condition  $M(s_i) = M(s_i^{-1})$  is automatically

satisfied if the number of elements in  $V$  is finite. Furthermore, our three conditions on the function  $\phi(x; s_j^e)$  for  $x_i \in V$ ,  $s_j$  a generator are:

$$(4.33) \quad \begin{aligned} &\phi(x; s_j^e) \text{ is an } x \in V \\ &\text{If } x; s_j^e \text{ is an } x \in V, \text{ then } \phi(x; s_j^e) = x; s_j^e \\ &\phi[\phi(x; s_j^e) s_j^{-e}] = x_i \end{aligned}$$

The second condition prescribes the value of  $\phi$  for certain arguments. If  $M(s_i) = M(s_i^{-1})$ , we may choose the rest of the values of  $\phi(x; s_i)$  to make  $x_j \rightarrow \phi(x; s_i)$  a one-to-one mapping of  $V$  onto itself and determine  $\phi(x; s_i^{-1})$  by the inverse mapping. Here all three conditions of (4.33) are satisfied, and it can be shown [15] that the elements of  $V$  are the coset representatives of the subgroup  $H$  generated by the  $u_{ij} = x; s_j \phi(x; s_j)^{-1}$  and the nontrivial  $u$ 's are free generators of  $H$ . From this the following theorem readily follows.

*Theorem 15.* Let  $F$  be a free group on generators  $s_i$ ,  $i \in I$  and  $w \neq 1$  any element of  $F$  not the identity. Then there is a normal subgroup  $K$  of finite index in  $F$  such that  $w \notin K$ .

PROOF. From the preceding we may take the word  $w$  and its initial segments as a Schreier system as coset representatives of a subgroup  $H$  of finite index and clearly  $w \notin H$ . If we represent  $F$  as a permutation group on the cosets of  $H$ , the kernel  $K$  of this representation is the desired normal subgroup.

*Corollary.* In a free group the intersection of all subgroups of finite index is the identity.

A group in which the intersection of all subgroups of finite index is the identity is said to be *residually finite*. In a residually finite group we may introduce a topology [11] calling a subgroup of finite index a neighborhood of the identity.

In a free group the Schreier method and the Nielsen method of finding free generators for a subgroup are related. If we take as representative of a coset the earliest elements in the alphabetical ordering the  $u_{ij}$ 's (or the inverses of some of them) are the generators yielded by the semialphabetical ordering. This has been shown in [15].

Easy consequences of the Nielsen method are the following theorems.

**Theorem 16.** *If a free group  $F$  of rank  $r$  is generated by  $r$  elements, it is freely generated by them, and mapping the generators of  $F$  onto these elements determines an automorphism of  $F$ .*

**Theorem 17.** *The automorphisms of a free group  $F$  on generators  $s_1, \dots, s_r$  are generated by the following automorphisms (1) permutations of the  $s_i$ ; (2)  $s_i \rightarrow s_i^{-1}$ ,  $s_j \rightarrow s_j$ ,  $j \neq i$ ; (3)  $s_i \rightarrow s_i s_j$  some  $j \neq i$ ,  $s_k \rightarrow s_k$ ,  $k \neq i$ .*

The first two types are *level* in the sense that the length of no element is changed by them. It is a curious thing that the length of an element in a free group is very important in most of our proofs, and yet it is not left invariant by most automorphisms. Indeed we do not know any good sense in which even on the average length is an invariant. Let us illustrate the difficulty. Let  $F$  be the free group with generators  $x, y$ . Then in  $F$  there are  $4 \cdot 3^{n-1}$  elements of (reduced) length  $n \geq 1$ . Consider the automorphism  $\alpha$  of  $F$  determined by  $\alpha: x \rightarrow xy, y \rightarrow y$ . Let  $T_n$  be the total length of the reduced forms  $(g)\alpha$  where  $g$  ranges over the  $4 \cdot 3^{n-1}$  elements of length  $n$ . Then it is not difficult to show that for  $n \geq 2$ ,  $T_{n+1} = 3T_n + 14 \cdot 3^{n-1}$  whence  $T_n = (14n + 4) \cdot 3^{n-2}$ ,  $n \geq 1$ . But the total of the original lengths  $L_n$  had the value  $4 \cdot 3^{n-1}n$ . Here  $T_n/L_n$  approaches the limit  $\frac{7}{8}$  as  $n$  goes to infinity, and so in this sense the automorphism is such that on the average the length of  $(g)\alpha$  is  $\frac{7}{8}$  times the length of  $g$ .

**Problem 5.** *Find a useful measure for elements, in a free group, related to length, which is invariant under automorphisms.*

The phrasing of this problem is of course vague since usefulness depends on the application desired. For example, suppose  $G$  is given by generators  $x_1, \dots, x_r$  and relations  $f_1(x) = \dots = f_m(x) = 1$ , that is, as the factor group of the free group  $F$  with generators  $s_1, \dots, s_r$  modulo the normal subgroup  $K$  generated by  $f_1(s), \dots, f_m(s)$  and their conjugates. It would be enormously valuable to have a measure on  $f_1(s), \dots, f_m(s)$  as elements of  $F$  which tells us whether or not  $G = F/K$  is finite and, if finite, what the order of  $G$  is. This is of course asking for a great deal. Nevertheless, in at least one very special case this is possible. Coxeter and Moser [p. 37] have shown that if  $G$  is generated by  $x$  and  $y$  subject to relations

$$x^u = y^v = (xy)^w = 1$$

then  $G$  is infinite if

$$(4.34) \quad \frac{1}{u} + \frac{1}{v} + \frac{1}{w} - 1 \leq 0$$

whereas the order of  $G$  is  $g$  if

$$(4.35) \quad 0 < \frac{1}{u} + \frac{1}{v} + \frac{1}{w} - 1 = \frac{2}{g}$$

The proof depends on a geometric argument involving the covering of a Euclidean, Riemannian, or Lobachevskian plane by triangles.

The most obvious invariant length for an element  $g$  of a free group  $F$  is the minimum length of  $(g)\alpha$  as  $\alpha$  ranges over all automorphisms of  $F$ . A procedure which determines this and considerably more is due to J. H. C. Whitehead [47]. The problem is to find a finite set  $T$  of automorphisms of  $F$  such that by successive applications of automorphisms of  $T$  to  $g$  we may reduce the length of the image  $(g)\alpha$  until no further reduction is possible. Of course the critical part of the argument is the proof that if  $L[(g)\alpha] \geq L(g)$  for every  $\alpha \in T$ , this is true for every automorphism  $\alpha$ . The original proof by Whitehead was entirely based on topological arguments. A direct algebraic proof was found by E. S. Rapaport [42], and a still better proof by P. Higgins and R. Lyndon [21].

**Theorem 18.** (Whitehead). *Let  $F$  be the free group generated by  $S = \{s_1, \dots, s_r\}$ . The level automorphisms of  $F$  are those given by mapping  $s_i \rightarrow s_j^{\epsilon_i(i)}$ ,  $\epsilon_i = \pm 1$ ,  $i = 1, \dots, r$  where  $i \rightarrow j(i)$  is a permutation. Let  $a$  be any one of the letters  $s_1, \dots, s_r, s_1^{-1}, \dots, s_r^{-1}$ . A Whitehead automorphism  $(A, a)$  of  $F$  maps  $a$  into  $a$  and each  $s_i \neq a, a^{-1}$  onto one of  $s_i, s_i a, a^{-1} s_i$ , or  $a^{-1} s_i a$ . Here  $A$  is the set of the letters consisting of  $a$  and those letters  $y \neq a, a^{-1}$  which are carried into either  $ya$  or  $a^{-1}ya$ , whence if  $x \rightarrow a^{-1}x$ , then  $x^{-1} \in A$ ,  $x \notin A$ . Let  $T$  be the set of level automorphisms and Whitehead automorphisms. Let  $g_1, g_2, \dots, g_m$  be a finite set of elements of  $F$ . If the total length of the  $g$ 's,  $L_S(g_1) + \dots + L_S(g_m)$  is not reduced by any automorphism of  $T$ , the total length of the  $g$ 's is not reduced by any automorphism of  $F$ , and an automorphism which leaves the total length unchanged is a level automorphism.*

This theorem gives an effective algorithm for finding a canonical form for a finite set of elements  $g_1, \dots, g_m$  under automorphisms

of  $F$ . We apply an automorphism of  $T$  if this reduces the total length of the  $g$ 's, and take the new set and apply the automorphisms of  $T$  to this. When no automorphism of  $T$  reduces the total length we are finished, and the result is unique up to a level automorphism.

It is clear that a set of automorphisms sufficient for the validity of the theorem must generate all automorphisms, since every nonlevel automorphism reduces the length of some element of  $F$ . Hence presumably the set should contain the generating automorphisms of Theorem 17. But these are not enough. If  $F$  is generated by  $x, y, z$ , the Whitehead automorphism  $\alpha: x \rightarrow x, y \rightarrow yx^{-1}, z \rightarrow xz$  reduces  $x^{-1}y^{-1}z^{-1}xyz$  to  $y^{-1}z^{-1}yz$  but one of the automorphisms of Theorem 17 reduces its length.

## 5. TARTAKOVSKII'S LIMITED SOLUTION OF THE WORD PROBLEM

In a series of three papers which appeared in 1949, V. A. Tartakovskii [45, 46] gave a method for solving the word problem for a certain class of finitely presented groups. It is to be presumed that his goal was the solution of the word problem for all finitely presented groups. Novikov's proof in 1955 that this is impossible by no means destroys the value of Tartakovskii's work. Indeed, Novikov's proof may be regarded as a justification for the special conditions imposed by Tartakovskii and raises the problem of widening these conditions.

Tartakovskii deals with relations imposed on the free product of cyclic groups, some of which may be of finite order, some of infinite order. But the sketch of his argument presented here is based on the presentation of a group as a factor group of a free group. In several respects this simplifies the line of argument, although the results are in some instances less general. We must give some notation and terminology before we can even formulate the results obtained.

Let  $G$  be a group presented as the factor group of the free group  $F$  generated by  $s_1, \dots, s_r$ , modulo the normal subgroup  $K$  of  $F$  generated by the finite set  $f_1(s), \dots, f_m(s)$  and their conjugates. This is equivalent to saying that  $G$  is generated by elements  $x_1, \dots, x_r$  subject to defining relations  $f_1(x) = 1, \dots, f_m(x) = 1$ . Our problem then is to develop an algorithm to decide whether or not a word  $g(x)$  is the identity, or alternately to decide whether or

not  $g(s)$  is in  $K$ , with the canonical homomorphism determined by  $s_i \rightarrow x_i$ ,  $i = 1, \dots, r$ .

If  $g = g_1 = a_1 a_2 \cdots a_n$  is a reduced word in the free group  $F$ , then  $a_1^{-1} g a_1 = a_2 \cdots a_n a_1$ ,  $(a_1 a_2)^{-1} g (a_1 a_2) = a_3 \cdots a_n a_1 a_2$ , and similarly  $g_i = a_i a_{i+1} \cdots a_n a_1 a_2 \cdots a_{i-1}$  for each  $i = 1, 2, \dots, n-1$  is a conjugate of  $g$ . Thus the conjugacy class of  $g$  as an element of  $F$  is the same as that of its cyclical rearrangements  $g_i$ ,  $i = 1, \dots, n$ . We say that  $g = a_1 a_2 \cdots a_n$  is *cyclically reduced* if  $g$  is reduced and also  $a_n \neq a_1^{-1}$ , in which case the cyclical rearrangements are also cyclically reduced. A word  $g = a_1 a_2 \cdots a_n$  in which  $a_1$  is considered as following  $a_n$  is called a *cyclic word*. If we wish to make the distinction we write  $g = a_1 a_2 \cdots a_n$  for the ordinary or *linear word* and  $g = [a_1 a_2 \cdots a_n]$  for the cyclic word. Here  $a_i a_{i+1} \cdots a_n a_1 a_2 \cdots a_{i-1}$  is the *split* of the cyclic word between  $a_{i-1}$  and  $a_i$ . And so we may think of the letters of a cyclic word as being on a circle to be read in a clockwise direction, and we obtain one of its linear forms by splitting between any two consecutive letters. It is easy to see that in a class of conjugate elements in the free group  $F$  there is a unique cyclically reduced cyclic word, as determined by any one of its linear splits.

Tartakovskii finds it convenient to use three terms for different kinds of equality. Let  $A = a_1 a_2 \cdots a_t$  and  $B = b_1 b_2 \cdots b_n$  be two words in  $F$ . We say that  $A$  and  $B$  are *graphically equal* (or *equal as written*), if  $t = n$ , and  $a_i = b_i$ ,  $i = 1, \dots, t$  and write  $A \approx B$ . If  $A$  and  $B$  are the same element of  $F$  (i.e., their reduced forms are equal), we say that  $A$  and  $B$  are *freely equal* and write  $A = B$ . We similarly speak of graphical and free equality for cyclic words. If  $AB^{-1} \in K$ , that is,  $A(x) = B(x)$  in  $G$ , we say that  $A$  and  $B$  are *equal in the group* and write  $A \equiv B$ .

If we split a cyclic word  $A$  at a certain point, a cyclic word  $B$  at a certain point, take the product of these linear words, and then form a cyclic word, we call this operation *composition* and write the result as  $A * B$ . If  $A = [a_1 a_2 \cdots a_t]$ ,  $B = [b_1 b_2 \cdots b_n]$ , then  $A * B = [a_i a_{i+1} \cdots a_t a_1 \cdots a_{i-1} b_j b_{j+1} \cdots b_n b_1 \cdots b_{j-1}]$  for appropriate  $i$  and  $j$ . We note that  $A * B = [x^{-1} A x \cdot y^{-1} B y]$  with  $x = a_1 \cdots a_{i-1}$ ,  $y = b_1 \cdots b_{j-1}$ . We can write  $A_{(i)} * B_{(j)}$  if we wish to indicate the points of splitting. It is of course true that the result of composition depends very much on the points of splitting, but for the most part we will not indicate these. Any equality which we write between compositions will mean that the

splitting points must be chosen appropriately. Composition is commutative:

$$(5.1) \quad A * B \approx B * A$$

This of course means that with the same splitting points the two are graphically the same as cyclic words. We have in addition a relation of *quasi-associativity*,

$$(5.2) \quad A * (B * C) \approx (A * B) * C \text{ or } (A * C) * B$$

Note that  $A * (B * C)$  is of one of the four forms (1)  $|A \ B \ C|$ , (2)  $|B \ A \ C|$ , (3)  $|B_1A \ B_2C|$  with  $B \approx B_1B_2$  or (4)  $|B \ C_1A \ C_2|$  with  $C \approx C_1C_2$ . With this observation the validity of (5.2) follows.

Tartakovskii solves the word problem for finitely presented groups  $G$ , where  $G = F/K$ ,  $F$  the free group generated by  $s_1, \dots, s_r$ , and  $K$  the normal subgroup generated by  $f_1(s), \dots, f_m(s)$  and their conjugates, providing the defining words  $f_1(s), \dots, f_m(s)$  do not combine too readily. This means that one or the other of two precise conditions holds. Without loss of generality we may take  $f_1, \dots, f_m$  to be cyclically reduced cyclic words  $A_1, \dots, A_m$ . If  $A = a_1a_2 \dots a_t$ , then  $A^{-1} = a_t^{-1} \dots a_1^{-1}$  and the composition  $A * A^{-1}$  where we divide at corresponding points is

$$(5.3) \quad |a_ia_{i+1} \dots a_t a_1 \dots a_{i-1} \cdot a_{i-1}^{-1} \dots a_1^{-1} a_t \dots a_i^{-1}| = 1$$

Of course if  $A$  is a power there may be more than one splitting with  $A * A^{-1}$  freely equal to the identity.

Excluding the compositions  $A * A^{-1}$  freely equal to the identity, let  $m_{ij}$  be the greatest length of  $A_i$  which is canceled in any composition  $A_i * A_j$  or  $A_i * A_j^{-1}$ . Further let  $n_i$  be the length of  $A_i$ . Define  $\delta$  by

$$(5.4) \quad \delta = \max \frac{m_{ij}}{n_i}, \quad \text{all } i, j$$

*Tartakovskii's First Condition.* The word problem is solvable for  $G$  if  $\delta < \frac{1}{8}$ .

Now suppose we can write  $A$  in the form

$$(5.5) \quad A = A^{(1)} A^{(2)} \dots A^{(k)}$$

where, excluding compositions  $A * A^{-1}$  freely equal to 1, there is a composition  $A * A_j$  or  $A * A_j^{-1}$  which completely cancels  $A^{(u)}$  for

an appropriate  $j$  depending on  $u$  for each  $u = 1, \dots, k$ . The minimum  $k$  for all  $A_i$ ,  $i = 1, \dots, m$ , and all such decompositions we shall call  $\bar{k}$ , the *index of erosion*.

*Tartakovskii's Second Condition.* The word problem is solvable for  $G$  if  $k > 6$ .

The first condition is really included in the second. There is simply a slight difference in the viewpoint. Both conditions are precise formulations of the loose assertion that the defining relations  $_j(x) = 1, j = 1, \dots, m$  do not combine too readily two at a time.

The major part of Tartakovskii's work does not depend on his conditions but is applicable to any finitely presented group. First we note that in general a few elementary reductions may be made. We may change the relations to equivalent ones so that each relation word  $A_1, \dots, A_m$  is cyclically reduced. Second, if some of the generators say  $s_{u+1}, \dots, s_r$  do not occur in any relation, then  $G$  is the free product of the group on  $s_1, \dots, s_u$  with the relations and the free group on  $s_{u+1}, \dots, s_r$ . Third, if a generator (or its inverse) occurs only once in a relation, we may use this relation to solve for this generator in terms of the rest and then have a group with one less generator and one less relation.

There are a number of theorems on compositions. We assume  $G$  to be given as  $F/K$  where  $s_1, \dots, s_r$  are the free generators and  $f_i(s) = A_i$ ,  $i = 1, \dots, m$  are cyclically reduced words which with their conjugates generate  $K$ , and that every  $s_j$  occurs in at least one  $A_i$ . Any word which is in  $K$  (i.e., is the identity in  $G$ ) is called a *Dyck word*.

*Theorem 19.* Every Dyck word is freely equal to a composition of the defining words  $f_i(s) = A_i$ ,  $i = 1, \dots, m$  and their inverses.

**PROOF.** If  $B$  and  $C$  are Dyck words so are their inverses, and their product  $BC$  is a composition. Also if  $B$  is a Dyck word and  $x$  is a generator, then by assumption there is an  $A_i$  which contains the generator  $x$  or  $x^{-1}$ . Hence  $A_i$  (or  $A_i^{-1}$ ) is cyclically equivalent to a word  $xT$ . But then the composition  $T^{-1}x^{-1}BxT$  is cyclically equivalent to  $x^{-1}BxTT^{-1}$ , and this is freely equal to  $x^{-1}Bx$ . This is sufficient to prove the theorem. Conversely, it is trivial that any word freely equal to a composition of the  $A_i$  and their inverses is a Dyck word. The foregoing theorem says that we may confine our attention to a study of compositions.

*Theorem 20. (Commutativity and Quasi-associativity)  $|A * B|$  and  $|B * A|$  are graphically equal.  $|A * (B * C)|$  is graphically equal to one of  $|(A * B) * C|$  or  $|(A * C) * B|$ .*

These results we have already proved.

*Theorem 21. (Monotone Compositions) Any composition of words  $A_1, A_2, \dots, A_n$  (not necessarily distinct) is graphically equal to a monotone composition*

$$|(((\dots (A_{i_1} * A_{i_2}) * A_{i_3}) * \dots) * A_{i_n})|$$

where  $i_1, i_2, \dots, i_n$  are  $1, \dots, n$  in some order and where we may choose the first factor  $A_{i_1}$  arbitrarily.

This theorem follows from the preceding without much difficulty. To simplify our notation we write  $A_{i_1} * A_{i_2} * \dots * A_{i_n}$  for the monotone composition of this theorem.

In a composition  $A_1 * \dots * A_n$ , a segment of consecutive letters coming from a factor  $A_i$  and bounded by letters from other factors is called an *arc*. As we construct  $A_1 * \dots * A_i$ ,  $i = 2, \dots, n$ , when we insert a new factor  $A_{i+1}$  the splitting point may be between two previous arcs, and the number of arcs is increased by one, but if the splitting point is in the interior of a previous arc we increase the number of arcs by two. This gives Theorem 22.

*Theorem 22. In a graphically free composition of  $n$  factors there are at most  $2n - 2$  arcs.*

Suppose in a composition  $P = A_1 * \dots * A_n$  two factors  $A_i, A_j$  are such that  $A_i A_j = 1$  and that in the composition  $P$  we have consecutive segments  $\dots A'_i M A'_j \dots$ , where  $M$  is a segment freely equal to the identity  $A'_i A'_j = 1$  and  $A_i, A_j$  are split so that  $A_i = A'_i A''_i$  and  $A_j = A'_j A''_j$  are graphically inverse. We say that the composition  $P$  is *complex*. A composition  $P$  is *simple* if it is not complex. Note that if  $A = aba^{-1}babB = A^{-1} = b^{-1}a^{-1}b^{-1}ab^{-1}a^{-1}$ , the split  $A'' = a^{-1}bab$ ,  $A' = ab$ ,  $B' = b^{-1}a^{-1}$ ,  $B'' = b^{-1}ab^{-1}a^{-1}$  does not make the composition complex since  $A''A'$  and  $B'B''$  are not graphical inverses.

A main feature of Tartakovskii's approach is to reduce his algorithms to an examination of simple compositions. In reducing a composition to a reduced cyclic word, there is no apparent way of estimating the amount of cancelation if this includes cancelation of

factors with parts of their own graphical inverses. But in a simple composition under appropriate conditions this is possible.

The next feature of Tartakovskii's attack is the theory of *extinction*. This is based on the alphabetical ordering of linear words introduced in Section 4. Tartakovskii orders  $s_i^{-1} < s_i^{-1}$  for a generator  $s_i$ , instead of  $s_i < s_i^{-1}$ , but this is an unimportant technical convention. The ordering of a word is the ordering of the reduced word to which it is freely equal.

If  $A > B$  and  $A$  is equal to  $B$  in the group  $G$ ,  $A \equiv B$  (i.e.,  $AB^{-1} \in K$  in  $F$ ) we say that  $A = B$  *extinguishes*  $A$ . The set of all extinguishable words is the *domain of extinction* of the group  $G$ , and the words which are not extinguishable are the *fundamental domain*. If we write

$$(5.6) \quad F = K + KX_2 + \cdots + KX_i + \cdots$$

where each coset representative  $x_i$  is the earliest element in its coset, the coset representatives  $x_i$  are the fundamental domain of  $G = F/K$  and all other words are in the domain of extinction. If for every word  $A$  in  $F$  we can find an earlier word  $B$  equal to  $A$  in  $G = F/K$  or show that  $A$  is in the fundamental domain, we shall have solved the word problem, since any word  $A$  has only a finite number of predecessors in the alphabetical ordering. This is what Tartakovskii does when one of his conditions holds.

Let  $P$  be a fixed reduced word. The set of all words  $xPy$  that are reduced as written is called a *ray*, and  $P$  is its *kernel*. If every word in  $xPy$  is extinguishable, this is called a *ray of extinction*. Trivially, if  $P$  is extinguishable, then  $xPy$  is a ray of extinction.

Let  $R$  be a linear Dyck word, but not necessarily reduced. Suppose  $R \approx PQ^{-1}$  where  $P > Q$ , that is,  $\bar{P} > \bar{Q}$  where  $\bar{P}, \bar{Q}$  are respectively the reduced forms of  $P$  and  $Q$ . Then  $x\bar{P}y$  is called a *left ray of extinction of  $R$* . The set of all left rays of extinction which can be formed from linear words of the cyclic word  $|R|$  is called the *left star of extinction of  $R$* . The *right star of extinction of  $R$*  is the left star of extinction of  $R^{-1}$  and the *star of extinction of  $R$*  is the sum of its left and right stars of extinction.

**Theorem 23.** *The domain of extinction of the group  $G$  is the sum of all the stars of extinction of all completely reduced Dyck words.*

The proof of this theorem is easy and is merely a formalization of the main problem; the real difficulties lie ahead.

Let  $B_1, B_2, \dots, B_t$  be a set of cyclic words not necessarily distinct nor reduced. Let

$$i_1 < i_2 < \dots < i_k, \quad k = 1, \dots, t$$

be an ordered subset of  $1, \dots, t$ , and let  $[B_{i_1}, B_{i_2}, \dots, B_{i_k}]$  be a subset (base) of  $B_1, \dots, B_t$ . Then the set of all the compositions which can be formed from the base will be designated by  $\{B_{i_1}, B_{i_2}, \dots, B_{i_k}\}$ . The sum of the stars of extinction of all compositions of the bases will be called the *zone of extinction* of the base  $B_1, \dots, B_t$  and written  $Z[B_1, \dots, B_t]$ .

The next theorem is a major one and requires a long sequence of preliminary lemmas of some difficulty. This is the longest and hardest part of the work.

*Theorem 24. The star of extinction of a complex composition is contained in the sum of the stars of extinction of simple compositions of fewer factors.*

If  $P = A_1 * \dots * A_n$  is complex with respect to  $A_i$  and  $A_j$ , it is shown that the star of extinction of  $P$  is contained in the zone of extinction  $Z[A_1, \dots, A_{i-1}, A_{i+1}, \dots, A_{j-1}, A_{j+1}, \dots, A_n]$ .

The next part of Tartakovskii's approach is a systematic way of finding the reduced form of a composition  $P = A_1 * \dots * A_n$ . Suppose there are  $N$  arcs in  $P$ . We know from Theorem 22 that  $N \leq 2n - 2$ . Starting with the first arc of  $A_1$ , let the arcs be called  $C_1, C_2, \dots, C_N$ , where of course as we move clockwise around the cyclic word,  $C_1$  follows upon  $C_N$ . Thus  $|P| = |C_1 C_2 C_3 \dots C_N|$ . Our first cancellation is to cancel as much of  $C_1$  as can be canceled by  $C_2$  and so if  $C_1 = C_{11} C_{12}$ ,  $C_2 = C_{12}^{-1} C_{22}$ ,  $P = |C_{11} C_{22} C_3 \dots C_N|$ . If  $C_{11}$  is the identity, we say that  $C_1$  has been annihilated. If  $C_{12}$  is the identity, there is no cancellation. Next we cancel as much of  $C_{22}$  as we can with  $C_3 C_{22} = C_{221} C_{222}$ ,  $C_3 = C_{222}^{-1} C_{32}$ ,  $P = |C_{11} C_{221} C_{32} C_4 \dots C_N|$ . In this manner we move clockwise around the circle as many times as necessary, canceling on the left as much as is possible from an arc which has not been annihilated. On the first circuit of the circle there are at most  $N$  of these cancellation steps, called reductions, since there are  $N$  arcs. After the first circuit no further cancellation is possible on the left between two arcs which were adjacent on a previous circuit; hence further cancellation is possible only when an arc has been annihilated. Since at worst all arcs are annihilated there are at most  $N$  reductions

possible beyond the first circuit. Let  $E$  be the total number of reductions. Thus we have shown the following theorem.

*Theorem 25.* If  $E$  is the total number of reductions required to put the composition  $P = A_1 * \cdots * A_n$  into reduced form, and  $P$  has  $N$  arcs, then  $E \leq 2N \leq 4n - 4$ .

In his third paper Tartakovskii is able to improve on this theorem in the following way.

*Theorem 26.* If  $P = A_1 * \cdots * A_n$ , there is a composition of  $A_1, \dots, A_n$  freely equal to  $P$  for which the total number of reductions  $E$  satisfies  $E \leq 3n - 2$ .

We have come this far without using the conditions at all. But now the application of the conditions is easy. We will use only the second condition, the first being a consequence of the second. Let  $P = B_1 * B_2 * \cdots * B_n$  be a simple composition of the defining words  $f_i(s) = A_i$ ,  $i = 1, \dots, m$ , and their inverses satisfying Theorem 26. Let us list  $B_1, B_2, \dots, B_n$  as cyclic words and at each reduction of  $P$  let us delete from the two corresponding  $B$ 's the letters which have been removed. We keep a record of the total number of reductions for each  $B$ , say  $b_i$  reductions applicable to  $B_i$ . Then at the end of all reductions we have

$$(5.7) \quad b_1 + b_2 + \cdots + b_n = 2E \leq 6n - 4$$

Since the composition was simple, at no stage of reduction were we canceling a segment of a  $B$  equal to  $A_i$  with the corresponding segment of a  $B$  equal to  $A_i^{-1}$ . By the second condition, at least seven such reductions are required to annihilate all of a  $B$ . Now for each  $i = 1, \dots, n$  let  $c_i = 0$  if  $b_i \geq 7$  and  $c_i = 7 - b_i$  if  $0 \leq b_i < 7$ . Then from (5.7) it readily follows that

$$(5.8) \quad c_1 + c_2 + \cdots + c_n \geq 7n - (6n - 4) = n + 4$$

Since the  $c$ 's are nonnegative and at most equal to 7, it follows that there are at least  $(n + 4)/7$  positive  $c$ 's in (5.8). But each positive  $c$  corresponds, by the second condition to a  $B$  which has not been completely canceled in the reduction and so its remaining length is at least 1. Hence  $D$ , the reduced form of  $P$ , has  $L_S(D) \geq (n + 4)/7$ . Usually we will have  $L_S(D) \geq n + 4$ , for if  $c_i$  is positive and if there is a decomposition (5.5) which includes the segments already reduced in  $B_i$ , the length of what remains must be at least  $c_i$ . If

this were valid for every  $i$ , we would have  $L_S(D) \geq n + 4$ . It might happen, however, that the only composition of  $B = A_i$  which cancels a particular letter is that which corresponds to cancellation with its own inverse. This situation arises, for example, if  $A_1 = abcabc$ ,  $A_2 = abab$  and applied to a letter  $c$  in  $A_1$ .

Hence we obtain all simple reduced Dyck words of length at most  $M$  by finding all simple compositions of at most  $7M - 4$  of the defining  $f_i(s) = A_i$ . If  $g(s)$  is of length  $n$ , then  $g(s)$  is in the fundamental domain of  $G$  unless it is in the star of extinction of some simple reduced Dyck word of length at most  $2n$ . Thus a listing of the simple reduced Dyck words of length at most  $2n$  enables us in a finite number of steps to decide whether or not  $g(s)$  is equal to the identity. Thus with Tartakovskii's conditions the word problem is solved.

*Problem 6. Find other conditions under which the Tartakovskii procedures yield a solution of the word problem.*

## 6. THE BURNSIDE PROBLEM

In 1902 Burnside [7] wrote, "A still undecided point in the theory of discontinuous groups is whether the order of a group may be not finite while the order of every operation it contains is finite." Burnside was of course discussing finitely generated groups, and his question can be expressed, "is a finitely generated periodic group necessarily finite?" The question in this form has not really been touched. If we assume the orders of the elements to be bounded and so all divisors of some integer  $n$ , the question takes the form which we now call the Burnside problem.

*The Burnside Problem. If a group  $G$  is finitely generated and if every element in  $G$  has finite order dividing an integer  $n$ , is  $G$  finite?*

If we let  $F_r$  be the free group with  $r$  generators  $s_1, \dots, s_r$  and  $N$  be the normal subgroup of  $F_r$  generated by all elements  $x^n$ ,  $x \in F_r$ , then  $B(n, r) = F_r/N$  is clearly a group generated by  $r$  elements in which  $Z^n = 1$  for every  $Z \in B(n, r)$  and so has the properties assumed in the problem. Clearly every group  $G$  generated by  $r$  elements in which  $Z^n = 1$  for every  $Z \in G$  is a homomorphic image of  $B(n, r)$  and so our problem may be put as "Are the groups  $B(n, r)$  finite?" It is known that the groups  $B(n, r)$  are finite when  $n = 2$ ,

3, 4, 6 [16, p. 320], and Novikov [40] has recently announced that the groups  $B(n, r)$  are infinite if  $n \geq 72$ . But Novikov's full proof has not as yet been published, and it is rumored that he may have to increase the number 72 to some larger number. Later in this section we shall discuss his announcement which depends on an analysis following the lines of Tartakovskii's work given in Section 5. But the groups  $B(n, r)$ , although finitely generated are not finitely related, and more seriously, the Tartakovskii conditions fail to hold.

A weaker form of the Burnside problem is known as the restricted Burnside problem: *For given integers  $n$  and  $r$  is there a number  $b_{n,r}$  such that every finite group with  $r$  generators of exponent  $n$  has order at most  $b_{n,r}$ ?*

Note that an affirmative answer to the Burnside problem is automatically an affirmative answer to the restricted problem. We also note that if  $F_r$  is the free group with  $r$  generators and if  $N_1$  and  $N_2$  are normal subgroups of finite index in  $F_r$  such that each of  $F_r/N_1$  and  $F_r/N_2$  is of exponent  $n$ , then this is also true for  $F_r/N$  with  $N = N_1 \cap N_2$ , since  $N$  is of finite index and contains all  $n$ th powers. Hence if the answer to the restricted Burnside problem is affirmative, there is a single largest finite group  $F_r/N$  with  $r$  generators of exponent  $n$  such that every other such group is a homomorphic image of it.

It has been shown by Kostrikin [27, 28] that the answer to the restricted Burnside problem is affirmative when  $n$  is a prime. This we shall discuss in a later section. We note that for values of  $n$  for which both Kostrikin's result and Novikov's are valid there is a largest finite group, but the Burnside group  $B(n, r)$  is infinite. This situation presents some complex implications and means in particular that a finitely generated group of prime exponent exists which has no subgroup of finite index.

For  $n = 2$  it has already been noted that

$$x^{-1}y^{-1}xy = (x^{-1}y^{-1}x)^2x^{-2}(xy)^2$$

and so  $B(2, r) = F_r/N_2$  is Abelian if  $N_2$  is the fully invariant subgroup of  $F_r$  generated by all squares. Hence  $B(2, r)$  is the elementary Abelian group of order  $2^r$ .

In his original paper Burnside proved the finiteness of the groups  $B(3, r)$  but their correct orders and precise description were given by Levi and van der Waerden [30], and may also be found in the writer's book [16, p. 320]. These results are expressed in terms

of commutators, where for arbitrary elements of a group  $G$  we write  $(x, y) = x^{-1}y^{-1}xy$ ,  $(x, y, z) = ((x, y), z)$ ,  $(x, y, z, w) = ((x, y, z), w)$ ,  $(x, y; z, w) = ((x, y), (z, w))$ . Then in a group  $G$  of exponent three the following relations on commutators hold:

$$\begin{aligned}
 (y, x) &= (x, y)^{-1} \\
 (x, y^{-1}) &= (x^{-1}, y) = (x, y)^{-1} \\
 (6.1) \quad (x, y, y) &= (x, y, x) = 1 \\
 (x, y, z) &= (y, z, x) = (y, x, z)^{-1} = (x, z, y)^{-1} \\
 (x, y, z, w) &= (x, y; z, w) = 1
 \end{aligned}$$

Of these relations the first is an identity in any group. The rest may be found directly by calculations involving two, three, and four generators. The last of these is easily derived from the others in the following way:

$$\begin{aligned}
 (x, y; z, w) &= (x, y, (z, w)) = ((z, w), x, y) = ((z, w, x), y) \\
 (6.2) \quad &= (x, z, w, y) = (x, z, y, w)^{-1} = ((x, y, z)^{-1}, w)^{-1} \\
 &= (x, y, z, w)
 \end{aligned}$$

But also,

$$(6.3) \quad (z, w; x, y) = ((x, y), z, w) = (x, y, z, w)$$

Hence  $(x, y; z, w) = (z, w; x, y) = (x, y; z, w)^{-1}$ . But as  $(x, y; z, w)^3 = 1$ , it follows that  $(x, y; z, w) = 1$  and with (6.2) the last relation is established.

It follows from the relations (6.1) that in  $B(3, r)$  generated by  $x_1, \dots, x_r$  every element  $g$  can be put in the form

$$(6.4) \quad g = x_1^{a_1} \cdots x_r^{a_r} (x_1, x_2)^{b_{12}} \cdots (x_i, x_j)^{b_{ij}} \cdots (x_i, x_j, x)^{c_{ijk}} \cdots$$

where for  $(x_i, x_j)$   $i < j$  and for  $(x_i, x_j, x_k)$   $i < j < k$  and the exponents are 0, 1, or 2. For  $g$  can be expressed by some word  $g = u_1 u_2 \cdots u_m$  where each  $u_i$  is an  $x_j^{\epsilon}$ , and using the identity  $wv = vw(w, v)$  we may first move all  $x_1$ 's to the left, then all  $x_2$ 's etc., until we have  $g$  as a product  $x_1^{a_1} \cdots x_r^{a_r} \cdot T$  where  $T$  is a product of commutators, and from the relations of (6.1) a commutator  $(w, v)$  is

either the identity or equal to one in (6.4) or its inverse, and the commutators themselves commute. If the elements of (6.4) for different exponents were not all different, then by putting appropriate  $x_j$ 's equal to 1 there would be such an equality involving only three generators, and direct calculation in  $B(3, 3)$  shows that this is not the case. The foregoing results may be stated as a theorem.

*Theorem 27. The group  $B(2, r)$  is the elementary Abelian group of order  $2^r$ . The group  $B(3, r)$  is of order  $3^{m(r)}$ ,  $m(r) = r + \binom{r}{2} + \binom{r}{3}$  and its elements have the canonical form of (6.4).*

In Burnside's original paper he proved that  $B(4, 2)$  is finite and of order at most  $2^{12}$ , and this is indeed its true order. Sanov [43] proved  $B(4, r)$  finite for all  $r$  by introducing certain intermediate groups. If  $B(4, r)$  is generated by  $x_1, \dots, x_r$ , let us define for every  $r$ ,  $H_{2r} = B(4, r)$  and  $H_{2r+1} = \{B(4, r), x_{r+1}^2\}$ . Thus for each  $i$ ,  $H_{i+1} = \{H_i, x\}$  with  $x^2 \in H_i$ , since for  $H_{2r+1}$ ,  $x = x_{r+1}^2$  and  $x^2 = 1$  and for  $H_{2r+2}$ ,  $x = x_{r+1}$ . We prove by induction on  $i$  that  $H_i$  is finite. Trivially  $H_2$  is of order 4.

Suppose that  $H = H_i$  is of finite order  $M_i$ . Then every element  $g$  of  $H_{i+1}$  can be put in the form

$$(6.5) \quad g = h_0 x h_1 x h_2 x \cdots x h_{n-1} x h_n$$

with  $h_j \in H = H_i$ , since  $x^2 \in H$ , and  $h_i \neq 1$ ,  $i = 1, \dots, n-1$ . There are  $M_i^2(M_i - 1)^{n-1}$  elements of this form, and so if we can find a bound on the  $n$  needed for all elements of  $H_{i+1}$  we will have shown that  $H_{i+1}$  is finite. This is what we shall do.

From the relation  $(hx)^4 = 1$  we have

$$(6.6) \quad \begin{aligned} xhx &= h^{-1}x^{-1}h^{-1}x^{-1}h^{-1} = h^{-1}x(x^2h^{-1}x^2)xh^{-1} \\ &= h^{-1}xh^*xh^{-1}, \quad h^* \in H \end{aligned}$$

For our purposes the form of  $h^*$  is not important. But applying (6.6) to  $xh_ix$  in (6.5) we have

$$(6.7) \quad g = h_0 x h_1 x \cdots x h_{i-1} h_i^{-1} x h_i^* x h_i^{-1} h_{i+1} x \cdots x h_n$$

Thus (6.7) is another expression for  $g$  with  $n$  unchanged,  $h_0, \dots, h_{i-2}$  unchanged, and  $h_{i-1}$  replaced by  $h_{i-1}h_i^{-1}$ . Thus if  $h_{i-1}h_i^{-1} = 1$  we may replace  $g$  by an expression of the form (6.5) with a smaller value for  $n$ . Repeating this procedure we may replace

$h_{i-2}$  by  $h_{i-2}(h_{i-1}h_i^{-1})^{-1}$  or  $h_{i-2}h_ih_{i-1}^{-1}$ , and so on. In this way we may replace  $h_1$  by any one of  $h_1, h_1h_2^{-1}, h_1h_3h_2^{-1}, \dots, h_1h_3 \dots h_{2s-1}h_{2s}^{-1} \dots h_4^{-1}h_2^{-1}, h_1h_3 \dots h_{2s+1}h_{2s}^{-1} \dots h_4^{-1}h_2^{-1}$ . If any one of these is 1 we may reduce the length of  $g$ . But with  $n \geq M_i + 1$ , either one of these is 1 or two of them are equal, say,

$$(6.8) \quad h_1h_3 \dots h_{2s-1}h_{2s}^{-1} \dots h_4^{-1}h_2^{-1} \\ = h_1h_3 \dots h_{2t-1}h_{2t}^{-1} \dots h_4^{-1}h_2^{-1}$$

with  $t > s$ . This implies

$$(6.9) \quad h_{2s+1} \dots h_{2t+1}h_{2t}^{-1} \dots h_{2s+2}^{-1} = 1$$

but this is an element which could have replaced  $h_{2s+1}$ . Similarly if the repetition involves  $h_1h_3 \dots h_{2s+1}h_{2s}^{-1} \dots h_4^{-1}h_2^{-1}$  we find that  $h_{2s+2}$  can be replaced by 1. Thus if  $n \geq M_i + 1$ , we can express  $g$  by a shorter expression. Hence any  $g$  can be expressed as a word with  $n \leq M_i$ . It easily follows that  $H_{i+1}$  is finite of order at most  $M_i^{M_i+1}$ . We have proved Theorem 28.

*Theorem 28. The groups  $B(4, r)$  are finite.*

The numerical bound for the order of  $B(4, r)$  implied by this proof is hopelessly unsatisfactory. Starting with  $2^2$  as the order of  $B(4, 1)$ , it gives  $2^{10240}$  as a limit on the order of  $B(4, 2)$  although the true value is  $2^{12}$ . C. R. B. Wright [49, 50] has found considerably more information about the groups  $B(4, r)$  which is discussed in Section 8.

For  $n = 6$  and any  $r$ , the answer to the Burnside problem is affirmative. Surprisingly enough the exact order of  $B(6, r)$  is known for every  $r$ , although the corresponding orders for  $B(4, r)$  are at present unknown. This depends on the deep results of Hall and Higman [19], which is discussed in Section 7.

In concluding this section some remarks will be made about Novikov's attack on the Burnside problem. This depends on the Morse-Hedlund sequence and some variant of the Tartakovskii solution of the word problem, this last not being described in sufficient detail so that the full proof can be reconstructed from Novikov's announcement.

The Morse-Hedlund sequence is an infinite sequence of 0's and 1's which begins

$$(6.10) \quad 0110100110010110 \dots$$

This may be described in various ways. If the sequence is

$$(6.11) \quad S = c_0 c_1 c_2 \cdots c_n \cdots$$

we have the following rule for  $c_n$ :

Express  $n$  in binary form. If the sum of the digits is even,  $c_n = 0$ , if odd,  $c_n = 1$ .

Alternately write  $a_0 = 0$ ,  $b_0 = 1$ , and define recursively  $a_1 = a_0 b_0$ ,  $b_1 = b_0 a_0$ ,  $\dots$ ,  $a_{m+1} = a_m b_m$ ,  $b_{m+1} = b_m a_m$ . Here  $a_m$  is a sequence of  $2^m$  digits, and as these are the first digits of  $a_{m+1}$ , it does not matter whether the initial digits are considered to be in  $a_m$  or any  $a_s$  with  $s > m$ . Thus we obtain an infinite sequence whose first  $2^m$  digits, for any  $m$ , are the digits of  $a_m$ , and this is the Morse-Hedlund sequence of (6.11). Morse and Hedlund [33] have proved that their sequence  $S$  has the property of Theorem 29.

*Theorem 29.*  $S$  does not contain a block of consecutive digits of the form  $EEe$  where  $E$  is a block and  $e$  is the first digit of  $E$ .

From (6.11) we form a sequence  $T$  of the letters  $a, b, c$

$$(6.12) \quad T = d_0 d_1 d_2 d_3 \cdots d_n \cdots$$

by the rules

$$(6.13) \quad \begin{aligned} d_n &= a \text{ if } c_n c_{n+1} = 01 \\ d_n &= b \text{ if } c_n c_{n+1} = 10 \\ d_n &= c \text{ if } c_n c_{n+1} = 00 \text{ or } 11 \end{aligned}$$

Thus (6.12) begins, using (6.10)

$$(6.14) \quad T = acbabcacbcabacb \dots$$

They prove Theorem 30.

*Theorem 30.*  $T$  does not have a block of letters of the form  $EE$ .

Now let  $G$  be a group of exponent  $n$  generated by elements  $a, b, c$ . Each initial segment of  $T$  yields an element of  $G$ . If these are all different, then  $G$  is infinite. If two different initial segments are the same element, then say  $W_1 = W_1 W_2$  and  $W_2 = 1$  is a Dyck word which is a block of  $T$  and by Theorem 30  $W_2$  contains no block of the form  $EE$ . Thus the crux of Novikov's proof, Novikov's

Announced Theorem, is this: for  $n \geq 72$  every Dyck word in  $B(n, r) = F_r/N$  contains a block of the form  $EE$ .

Thus for  $r \geq 3$  the sequence  $T$  shows  $B(n, r)$  to be infinite. For  $r = 2$  we may replace  $a, b, c$  by  $1, 2, 3$  and use them as exponents in  $x^1y^3x^2y^1 \dots$ , and again conclude that  $B(n, 2)$  is infinite. The theorem depends on showing that in the decompositions (5.5) relevant to Tartakovskii's second condition there is always a segment  $A^{(j)}$  remaining after the Dyck word has been reduced which contains a block of the form  $EE$ . Thus Novikov's analysis must be comparable to Tartakovskii's, but the objective is different, namely to prove the existence of the block  $EE$  rather than find an estimate on length.

The role of the Morse-Hedlund sequence may be replaced by the following sequence recently found by R. Dean [9]:

$$D = xyx^{-1}y^{-1}xy^{-1}x^{-1}y \dots$$

The rule for construction of Dean's sequence  $D$  is the following. Let 1 stand for  $x$ , 2 for  $y$ , 3 for  $x^{-1}$ , 4 for  $y^{-1}$ . Begin with 1 2 3 4. At stage  $i$  divide all the numbers into 4 blocks of equal length  $A_1A_2A_3A_4$ . Stage  $i + 1$  consists of  $A_1A_2A_3A_4A_1A_4A_3A_2$ . Iteration yields an infinite sequence  $x^1y^2x^3y^4 \dots$ , with no consecutive block repeat.

## 7. THEOREMS ON $p$ -LENGTH. THE BURNSIDE PROBLEM FOR $n = 6$

A finite group  $G$  is said to be  $p$ -solvable,  $p$  a prime, if there is a chain of normal subgroups such that every factor group is either a  $p$ -group or has order prime to  $p$ , which latter we shall call a  $p'$ -group or non- $p$ -group. Clearly a solvable group is  $p$ -solvable for every prime  $p$ . A  $p$ -solvable group has a unique upper  $p$ -series

$$(7.1) \quad 1 \subseteq N_0 \subset P_1 \subset N_1 \cdots \subset P_i \subset N_i \subset P_{i+1} \cdots \\ \qquad \qquad \qquad \subset P_l \subseteq N_l = G$$

where  $N_0$  (possibly 1) is the largest normal non- $p$ -subgroup,  $P_1/N_0$  is the largest normal  $p$ -group in  $G/N_0$ , and generally  $N_i/P_i$  is the largest normal non- $p$ -subgroup in  $G/P_i$ , and  $P_{i+1}/N_i$  is the largest normal  $p$ -subgroup in  $G/N_i$ . Here we call the least  $l$  for which  $N_l = G$  the  $p$ -length of  $G$  and write  $l = l_p = l_p(G)$ . The concept of  $p$ -length was introduced in a major paper by Philip Hall and

Graham Higman [19], and its significance and relationship to the Burnside problem were developed there. They begin with a fundamental lemma.

*Lemma 5.* For all  $i \geq 1$ ,  $P_i$  contains the centralizer of  $P_i/N_{i-1}$  and  $N_i$  contains the centralizer of  $N_i/P_i$ .

An improvement of this for the  $p$ -factors is the following:

*Lemma 6.* If  $F_i/N_{i-1}$  is the Frattini subgroup of the  $p$ -group  $P_i/N_{i-1}$ , then  $P_i$  is the centralizer of  $P_i/F_i$ .

The force of this lemma is that  $G/P_i$  is faithfully represented by automorphisms of  $P_i/F_i$  induced by conjugation. Since  $V = P_i/F_i$  is an elementary Abelian  $p$ -group, if we write  $V$  additively the automorphisms of  $V$  are nonsingular linear transformations of  $V$  regarded as a vector space over the finite field  $GF(p)$ . This is an important instance of *internal representation* in which the normalizer  $N(A)$  of a subgroup  $A$  is represented by the automorphisms induced in  $A$ . The study of internal representations has in the last few years been of the greatest value, and goes well beyond the scope of this monograph.

Here  $H = G/P_1$  is a  $p$ -solvable group which has no normal  $p$ -subgroup and is faithfully represented on the elementary Abelian vector space  $V = P_1/F_1$ .

The object of the paper is to relate the  $p$ -length of  $G$  to various properties of a Sylow  $p$ -subgroup  $S(p)$  of  $G$ . In general, the greater  $\ell(p)$  is, the larger and more complex  $s(p)$  must be. They relate  $\ell(p)$  to the quantities (1)  $b = b_p$  where  $p^b$  is the order of  $S(p)$ , (2)  $c = c_p$  the class of  $S(p)$ , (3)  $d = d_p$  the length of the derived series of  $S(p)$ , and (4)  $e = e_p$  where  $p^e$  is the exponent of  $S(p)$ , that is the least  $e$  for which  $x^{p^e} = 1$  for every  $x \in S(p)$ . For application to the Burnside problem the exponent  $e_p$  is the most significant of these properties.

Their first main theorem is Theorem 31.

*Theorem 31.* Let  $H$  be a  $p$ -solvable linear group over a field of characteristic  $p$ , with no normal  $p$ -subgroup greater than 1. If  $g$  is an element of order  $p^n$  in  $H$  then the minimal equation of  $g$  is  $(x - 1)^r = 0$  where (1) if  $p$  is neither 2 nor a Fermat prime,  $r = p^n$ ; (2) if  $p$  is an odd Fermat prime  $r = p^n$  if  $H$  has Abelian Sylow 2-subgroups and  $p^{n-1}(p - 1) \leq r \leq p^n$  in any case; (3) if  $p = 2$ ,  $r = p^n$  if  $H$  has

*Abelian Sylow  $q$ -subgroups for all Mersenne primes  $q$  less than  $2^n$ ; if  $q = 2^s - 1$  is the least Mersenne prime for which this is not so,  $2^{n-s}q \leq r \leq 2^n$  and in any case  $2^{n-2} \cdot 3 \leq r$ .*

These results are best possible, and the way in which the Mersenne and Fermat primes enter in shows the deep arithmetical side of these results. These lead in turn to Theorem 32.

*Theorem 32. If  $G$  is a  $p$ -solvable group, where  $p$  is an odd prime, then (1)  $d_p \geq \ell_p$ , and (2)  $e_p \geq \ell$  if  $p$  is not a Fermat prime, and  $e_p \geq [\frac{1}{2}(\ell + 1)]$  if  $p$  is a Fermat prime. Furthermore, these results are best possible.*

There are also similar relations involving  $b_p$  and  $c_p$ . For  $p = 2$  the situation is a little more complicated. It has been shown by A. H. M. Hoare [25] that for  $p = 2$   $\ell_p \leq 3e_p$  and more recently by F. Gross [10], that  $\ell_p \leq 2e_p$ , although it may be true that  $\ell_p \leq e_p$ .

From Theorem 32, Hall and Higman are able to attack the restricted Burnside problem for solvable groups.

*Problem  $S_n$ . Does there exist for each positive integer  $k$  an integer  $s(n, k)$  such that every finite solvable group of exponent  $n$  that can be generated by  $k$  elements has order at most  $s(n, k)$ ?*

Let us remark that a solvable group  $G$  which is finitely generated and is of exponent  $n$  is certainly finite. For  $G/G'$  is a finitely generated Abelian group of exponent  $n$  and so is finite. By Theorem 13,  $G'$  is finitely generated and so  $G'/G''$  is a finite group. Continuing each derived factor  $G^{(i)}/G^{(i+1)}$  is finite and since  $G$  is solvable there is an integer  $t$  such that  $G^{(t)} = 1$ . Thus in  $G \supset G' \supset G'' \supset \dots \supset G^{(t)} = 1$ , each factor group is finite and so  $G$  itself is finite.

*Theorem 33. If  $n = p_1^{a_1} p_2^{a_2} \dots p_r^{a_r}$  is the factorization of the integer  $n$  as a product of prime powers  $q_i = p_i^{a_i}$ ,  $i = 1, \dots, r$ , then the answer to the problem  $S_n$  is affirmative if the answer to each  $S_{q_i}$ ,  $i = 1, \dots, r$  is affirmative.*

PROOF. We use induction on  $r$ , and take  $p_1 = 2$  if  $n$  is even. Then  $n = n_1 q_r$  with  $n_1 = p_1^{a_1} \dots p_{r-1}^{a_{r-1}}$  and  $q_r = p_r^{a_r}$ . Here  $q_r$  is odd, and by Theorem 32 if  $G$  is any finite solvable group generated by  $k$  elements of exponent  $n$ , then considering  $G$  as a  $p_r$ -solvable group there is a bound on  $\ell_{p_r} = \ell$  for the series (7.1) for  $G$ . But then

$G = N_\ell$  and  $N_\ell/P_\ell$  is a group of exponent  $n_1$  generated by  $k$  elements and so its order is bounded by  $s(n_1, k)$ . Then by Theorem 13  $P_\ell$  is finitely generated by say  $k_1$  elements, and since  $P_\ell/N_{\ell-1}$  is a finitely generated group of exponent  $q_r$ , its order is bounded by  $s(q_r, k_1)$  whence by Theorem 13  $N_{\ell-1}$  is finitely generated. Thus in turn we may find a bound on the order of each factor group in the series (7.1), and as the number of factors is also bounded, by Theorem 32 it follows that there is a bound on the order of  $G$ . Thus an affirmative answer to  $S_n$  is a consequence of an affirmative answer to  $S_{n_1}$  and  $S_q$ , and our theorem is proved.

For  $n = 6$  the known finiteness of  $B(2, r)$  and  $B(3, r)$  for all  $r$  gives an affirmative answer to Problem  $S_6$ . From Burnside's celebrated result that groups of order  $p^a q^b$ ,  $p$  and  $q$  primes, are solvable, it follows that a finite group of exponent 6, being of order  $2^a 3^b$ , is necessarily solvable. Thus  $S_6$  is equivalent to the restricted Burnside problem for  $n = 6$ , and we have an affirmative answer to it. We may even evaluate  $s(6, k)$ . Let  $G$  be finite of exponent 6 and let (7.1) be the upper 2-series for  $G$ . Then a Sylow 2-subgroup is elementary Abelian by Theorem 27 and so centralizes any part of it. Thus by Lemma 5 as  $P_1/N_0$  is a 2-group,  $P_1$  must contain an entire Sylow 2-subgroup, and so  $\ell_2(G) = 1$ . Using (2) in Theorem 31 it follows that  $\ell_3(G) = 1$ . This may also be proved directly without an enormous amount of labor. Thus there are two normal series for  $G$ .

$$(7.2) \quad \begin{aligned} 1 &\subseteq A_1 \subset A_2 \subseteq A_3 = G \\ 1 &\subseteq B_1 \subset B_2 \subset B_3 = G \end{aligned}$$

where  $A_1, A_3/A_2, B_2/B_1$  are 3-groups and  $B_1, B_3/B_2, A_2/A_1$  are 2-groups, and so  $A_2/A_1$  is a Sylow 2-subgroup of  $G$  and  $B_2/B_1$  is a Sylow 3-subgroup of  $G$ .

Now if  $G$  is generated by  $k$  elements,  $G/A_2$  is of exponent 3 and so of order at most  $3^{m(k)}$ ,  $m(k) = k + \binom{k}{2} + \binom{k}{3}$  by Theorem 27, whence by Theorem 13,  $A_2$  has  $a = 1 + (k-1)3^{m(k)}$  generators, and so  $A_2/A_1$  is of order at most  $2^a$ . Similarly  $G/B_2$  is of exponent 2 and of order at most  $2^k$ , whence  $B_2$  has  $b = 1 + (k-1)2^k$  generators, and consequently  $B_2/B_1$  being of exponent 3 has order at most  $2^{b + \binom{b}{2} + \binom{b}{3}}$ . These give upper bounds on the order of a Sylow 2-subgroup and a Sylow 3-subgroup, and so we have a bound on the

order of  $G$ ,  $s(6, k)$

$$(7.3) \quad s(6, k) = 2^a 3^{b+ \binom{b}{2} + \binom{b}{3}}, \quad a = 1 + (k-1)3^{k+ \binom{k}{2} + \binom{k}{3}}, \\ b = 1 + (k-1)2^k$$

Theorem 33 also gives an affirmative answer to  $S_{12}$  since this is known for  $S_3$  and  $S_4$ , but in the absence of satisfactory values for  $s(4, k)$  we are worse off in attempts to estimate  $s(12, k)$ .

Using the Hall-Higman results as motivation, the writer [14] has solved the Burnside problem for exponent 6.

*Theorem 34. The answer to the Burnside problem for exponent 6 is affirmative. The group  $B(6, k)$  is of order  $2^a 3^{b+ \binom{b}{2} + \binom{b}{3}}$ ,  $a = 1 + (k-1)3^{k+ \binom{k}{2} + \binom{k}{3}}$ ,  $b = 1 + (k-1)2^k$ .*

If  $F_k$  is the free group with  $k$  generators and  $N_2, N_3$  respectively the subgroups generated by squares and cubes, then  $F_k/N_2$  is of order  $2^k$ ,  $F_k/N_3$  is of order  $3^{k+ \binom{k}{2} + \binom{k}{3}}$  and  $N_2$ , and by Theorem 13,  $N_3$  has  $a$  generators and  $N_2$  has  $b$  generators. Then if  $N_2(N_3)$  is generated by the squares in  $N_3$  and  $N_3(N_2)$  is generated by the cubes in  $N_2$ , then  $N_3/N_2(N_3)$  is of order  $2^a$  and  $N_2/N_3(N_2)$  is of order  $3^{b+ \binom{b}{2} + \binom{b}{3}}$ . Now for any  $x$  in  $F_k$ ,  $x^6 = (x^3)^2 = (x^2)^3$  and so  $x^6 \in N_2(N_3)$  and  $x^6 \in N_3(N_2)$ . Hence  $N_6$ , the subgroup of  $F_k$  generated by sixth powers, is contained in  $W = N_2(N_3) \cap N_3(N_2)$ . Thus  $N_6 \subseteq W$ , and  $F_k/W$  is finite and has its order divisible by the number in the theorem. This number being the same as  $s(6, k)$  given in (7.3), is the order of  $F_k/W$ , but it is still conceivable that  $N_6 \subset W$  and  $F_k/N_6$  is infinite. Hence if we can show  $B(6, k)$  to be finite, we know what its order must be.

We assume that  $G = B(6, k) = F_k/N_6$  is generated by  $k$  elements and is of exponent 6. Then there is a series

$$(7.4) \quad G \supset M \supset M'$$

where  $M$  is generated by cubes of elements of  $G$  and so  $G/M$  is a 3-group and indeed  $G/M = B(3, k)$  is of order  $3^{m(k)}$ ,  $m(k) = k + \binom{k}{2} + \binom{k}{3}$  and so  $M$  has  $a = 1 + (k-1)3^{m(k)}$  generators.

Group  $M/M'$  is a finite group of exponent 2 and order  $2^a$ . Group  $M'$  is finitely generated;  $M$  is generated by cubes and is also finitely generated, and so  $M$  is generated by the finite number of cubes

needed to express this finite number of generators. If  $x_1, \dots, x_t$  are the cubes needed to generate  $M$ , they are of course elements of order 2.  $M'$  is generated finitely and also by  $x_i^{-1}x_j^{-1}x_ix_j$ ,  $i, j = 1, \dots, t$  and their conjugates and so by a finite number of these conjugates. Hence  $M'$  is generated by a finite number of elements of the form  $abab$ , where  $a^2 = 1$ ,  $b^2 = 1$ . We now turn to a lemma.

*Lemma 7. If a group  $H$  is generated by elements  $v_1, \dots, v_n$  and if any four  $v$ 's generate a group of exponent 3, then  $H$  is of exponent 3.*

This is an easy lemma depending on the observation that the Levi-van der Waerden relations in equation (6.1) involve at most four different generators. We now need to prove the following:

*Lemma 8. If  $a^2 = b^2 = c^2 = d^2 = e^2 = f^2 = g^2 = h^2 = 1$  in a group of exponent 6, then the subgroup  $K = \{abab, cdcd, efef, ghgh\}$  is of exponent 3.*

Once Lemma 8 is proved, we may use Lemma 7 to conclude that  $M'$  is of exponent 3 and so finite, and consequently that  $G = B(6, k)$  is finite, thus proving our theorem. The major part of the paper consists of calculations needed to establish Lemma 8. Lemma 8 might conceivably be put onto a high-speed computer, since it is an explicit, although reasonably difficult problem. It is interesting to see how the general problem as to the finiteness of  $B(6, k)$  can be reduced to the very explicit problem of proving Lemma 8.

*Problem 7. Find more accurate descriptions of the groups  $B(4, k)$  and their orders.*

*Problem 8. Is there a definite integer  $m$  such that if  $H$  is generated by  $x_1, \dots, x_n$  for any  $n$ , then  $H$  is of exponent 4 if any  $m$  of the  $x$ 's generate a group of exponent 4?*

*Problem 9. Prove the groups  $B(12, k)$  are finite using the finiteness of all  $B(2, r)$ ,  $B(3, s)$ ,  $B(6, t)$  for any  $r, s, t$ .*

The 3-length of a group of exponent 12 is at most two, and it has recently been shown by F. Gross that the 2-length is at most two.

## 8. THE RESTRICTED BURNSIDE PROBLEM. COMMUTATORS AND LIE RINGS

We use the notations  $(x, y) = x^{-1}y^{-1}xy$  for the commutator of elements  $x$  and  $y$  in a group  $G$ . When  $H$  and  $K$  are subgroups of  $G$

we write  $(H, K)$  for the subgroup generated by all commutators  $(h, k)h \in H, k \in K$ . Given a group  $G$  the subgroups  $G_1 = G$ , and recursively  $G_i = (G_{i-1}, G)$ ,  $i = 2, 3, \dots$ , form the *lower central series* of  $G$ .

$$(8.1) \quad G \supseteq G_2 \supseteq G_3 \supseteq \dots \supseteq G_n \supseteq G_{n+1} \supseteq \dots$$

From the definition of the lower central series it follows that if for some  $n$  we have  $G_n = G_{n+1}$ , then

$$(8.2) \quad G_n = G_{n+1} = G_{n+2} = \dots = G_{n+i} = \dots$$

A group  $G$  is *nilpotent* if there is an  $N$  such that  $G_N = 1$ . A finite group  $G$  is known to be nilpotent if and only if it is a direct product of  $p$ -groups, which are of course its Sylow subgroups. Thus a group of exponent  $p^s$ ,  $p$  a prime, is nilpotent if it is finite. If  $G$  is a free group with  $r$  generators  $G_n/G_{n+1}$  is a free Abelian group with  $\mu_r(n)$  generators where

$$(8.3) \quad \mu_r(n) = \frac{1}{n} \sum_{d|n} \mu\left(\frac{n}{d}\right) r^d$$

Here  $\mu(m)$  is the Möbius function. This formula was found by Witt [48] and may be found in the writer's book [16, p. 171]. Thus if  $G$  is generated by  $r$  elements and is of exponent  $p^s$ , each of the factor groups  $G_i/G_{i+1}$  is finite and so  $G$  will be finite if and only if  $G$  is nilpotent. Furthermore, if  $G_N = 1$  for some large  $N$ , and if we know that  $G_n = G_{n+1}$  for some  $n$ , it follows from (8.2) that  $G_n = 1$ . Hence the restricted Burnside problem is answered in the affirmative if there exists an  $n$  such that  $G_n = G_{n+1}$ . It was shown by Kostrikin [26] that if  $G = B(5, 2)$ , then  $G_{13} = G_{14}$  and that the largest finite group of exponent 5 with two generators is of order  $5^{33}$  (or  $5^{34}$  depending on a remaining ambiguity in his calculations). Graham Higman [22] solved the restricted Burnside problem for exponent 5 and any number of generators. Kostrikin [28] solved the problem for prime exponent  $p$ .

Let us represent conjugation by an exponent, writing  $x^{-1}ax = a^x$ . With this notation and the convention already introduced  $((x, y), z) = (x, y, z)$  there are a number of identities on commutators:

$$(8.4) \quad (x, x) = 1, (y, x) = (x, y)^{-1}$$

$$(8.4a) \quad (x, y, z) = (x, z)^y(y, z) = (x, z)(x, z, y)(y, z)$$

$$(8.5) \quad (x, yz) = (x, z)(x, y)^z = (x, z)(x, y)(x, y, z)$$

$$(8.6) \quad (x, y^{-1}, z)^y(y, z^{-1}, x)^z(z, x^{-1}, y)^x = 1$$

$$(8.7) \quad (x, y, z)(y, z, x)(z, x, y) \\ = (y, x)(z, x)(z, y)^x(x, y)(x, z)^y(y, z)^x(x, z)(z, x)^y$$

It is known that  $(G_i, G_j) \subseteq G_{i+j}$ , and it readily follows that the right-hand side of (8.7) belongs to  $(G_2, G_2) \subseteq G_4$ .

The identities (8.4)–(8.7) resemble very closely those for a Lie ring. If  $R$  is an associative ring we may define a Lie product  $[x, y]$  in  $R$  by the rule

$$(8.8) \quad [x, y] = xy - yx$$

With the rule (8.8) and the notation  $[[x, y], z] = [x, y, z]$ , we readily verify

$$(8.9) \quad [x, x] = 0, \quad [y, x] = -[x, y]$$

$$(8.10) \quad [x + y, z] = [x, z] + [y, z]$$

$$(8.11) \quad [x, y + z] = [x, y] + [x, z]$$

$$(8.12) \quad [x, y, z] + [y, z, x] + [z, x, y] = 0$$

Of these (8.11) is known as the Jacobi identity. It has been shown by Witt [48] and by Garrett Birkhoff [4] that a ring  $L$  with an addition, which is an Abelian group, and a bracket product  $[x, y]$  satisfying the laws of (8.9)–(8.12) can be realized in an appropriate associative ring  $R$  with the rule (8.8).

Following Graham Higman [23], let us show how to derive a Lie ring  $L$  from a group  $G$ . We choose a series of normal subgroups of  $G$ .

$$(8.13) \quad G = H_1 \supseteq H_2 \supseteq H_3 \cdots \supseteq H_i \supseteq \cdots$$

with the property that for all integers  $i, j$ , we have  $(H_i, H_j) \subseteq H_{i+j}$ . (Clearly the lower central series (8.1) is a special case of this.) The additive group of  $L$  is the direct sum of the (necessarily Abelian) factor groups  $H_i/H_{i+1}$  of the series. We call elements of  $H_i/H_{i+1}$  homogeneous of degree  $i$ . Bracket multiplication of homogeneous elements is defined by

$$(8.14) \quad [g_i H_{i+1}, g_j H_{j+1}] = g_i^{-1} g_j^{-1} g_i g_j H_{i+j+1}$$

This equation may be extended by the bilinearity rules (8.10) and

(8.11) to the whole of  $L$ . The rule (8.7) in  $G$  gives the Jacobi identity (8.12) for homogeneous elements in  $L$ , and bilinearity extends this to arbitrary element  $s$  of  $L$ . For a detailed treatment of the relation between commutators in groups and Lie rings see M. Lazard [29]. This correspondence gives information bearing on all of  $G$  only if for  $K = \bigcap_{i=1}^{\infty} H_i$  we have  $K = 1$ , and otherwise gives information on  $G/K$ .

For the Burnside problem the relationship between commutators and Lie rings is useful only when the exponent is a prime power  $p^s$  and here only for the restricted Burnside problem. If  $G$  is of exponent  $p^s$ , generated by  $k$  elements, and if  $L$  is the Lie ring associated with  $G$  by the rule (8.14), then  $L^n = 0$  becomes  $G_n = G_{n+1}$ , and by (8.2) we have  $G_n = 1$ . With the same  $n$  for every such finite  $G$  we know there is a bound  $b(p^s, k)$  on the order of  $G$ , and we have an affirmative answer to the restricted Burnside problem. Thus the solution of the restricted Burnside problem for prime power exponents is equivalent to showing that the associated Lie ring  $L$  is nilpotent.

Kostrikin [27, 28] solved the restricted Burnside problem for prime exponent  $p$ , which begins with a result by Philip Hall [18]. This is the result

$$(8.15) \quad (y, x, \overbrace{\dots}^{p-1}, x) = 1 \pmod{G_{p+1}}$$

for any  $x$  and  $y$  in a group  $G$  of exponent  $p$ . Here the commutator on the left is of weight  $p$ . In the associated Lie ring this becomes

$$(8.16) \quad [y, x, \dots, x] = 0 \quad p-1, x's$$

In (8.15) we are using the notation  $((a_1, \dots, a_{n-1}), a_n) = (a_1, \dots, a_n)$  and similarly in (8.16). The condition (8.16) is called the Engel condition of order  $p-1$ . To obtain (8.16) from (8.15) for all elements of  $L$  and not merely homogeneous ones requires a little care. For the proof of this see Graham Higman [22]. Kostrikin shows that a Lie ring  $L$  of characteristic  $p$  satisfying the  $m$ th Engel condition with  $m < p$  always contains a nilpotent ideal. From the general theory of Lie rings it follows that a finitely generated  $L$  with this property is necessarily nilpotent and so for an appropriate  $n$  we have  $L^n = 0$  whence  $G_n = G_{n+1}$ , and we have an affirmative answer to the restricted Burnside problem.

Kostrikin assumes that  $L$  does not contain a nilpotent ideal, and then finds a nilpotent ideal in  $L$ . This contradiction proves his result. For a Lie ring  $L$  we define a ring  $D(L)$ , the ring of inner derivations, in the following way. For  $x \in L$  let  $A(x)$  be the mapping of  $L$  into itself

$$(8.17) \quad A(x): u \rightarrow [u, x] \quad \text{all } u \in L$$

Here  $A(x)$  is an endomorphism of the additive group of  $L$ . The ring  $D(L)$  is the ring generated by the  $A(x)$  and is of course associative. From the rules (8.9)–(8.12) it follows that

$$(8.18) \quad A(x + y) = A(x) + A(y)$$

$$(8.19) \quad A([x, y]) = A(x)A(y) - A(y)A(x) = [A(x), A(y)]$$

Here (8.19) depends chiefly on the Jacobi identity (8.12) and the brackets on the right in (8.19) are those given by the rule (8.8) in the associative ring  $D(L)$ . If  $L$  contains no nilpotent ideal, as Kostrikin's assumes, the mapping (8.17) of  $L$  into  $D(L)$  is faithful. For simplicity Kostrikin writes  $x$  for  $A(x)$  as he has assumed the mapping to be faithful. Thus the Engel condition (8.16) takes the form in  $D(L)$

$$(8.20) \quad x^{p-1} = 0$$

The condition  $a_1 a_2 \cdots a_m = 0$  in  $D(L)$  is equivalent to  $[u, a_1, a_2, \dots, a_m] = 0$  in  $L$  for all  $u$ .

The main line of Kostrikin's argument can be expressed by several lemmas.

*Definition.* Let  $c_m$  denote an element of  $D(L)$  not zero such that  $c_m b^i c_m = 0$ ,  $i = 0, \dots, 2m - 1$ , for any  $b$  in  $D(L)$ .

*Lemma 9.* If there is a  $c_m$ ,  $m \geq 2$ , then one of the following elements is a  $c_{m+1}$ :

$$c_m, c'_m = [c_m a^{2m+1} c_m] \text{ for some } a$$

$$\text{or,} \quad [c'_m a^{2m+1} c'_m] \text{ for some } a \text{ and } d$$

*Lemma 10.* If  $L$  satisfies the  $n$ th Engel condition  $n < p$  and if  $m_0$  is the greatest integer not exceeding  $(n - 1)/2$ , the ideal  $N$  generated by a  $c_{m_0}$  is nilpotent, and indeed  $N^2 = 0$ .

The combination of these yields the following:

*Lemma 11.* If  $D(L)$  contains an element  $c_2$  then  $L$  contains a nilpotent ideal. The rest of the work consists in finding an element  $c_2$ .

*Lemma 12.*  $D(L)$  contains an element  $c \neq 0$  with  $c^2 = 0$ .

*Lemma 13.*  $D(L)$  contains an element  $c_2$ .

Kostrikin begins by remarking that if  $v^m = 0$  in  $D(L)$  and if  $v^{m-1} \neq 0$ , then for any  $w$  in  $D(L)$   $[wv^{m-1}]^{m-1} = 0$  providing that  $4 \leq m$ . This gives him an element  $b \neq 0$  with  $b^3 = 0$ . The rest of the argument involves some complicated calculation. This gives his main result, since he has a contradiction unless  $L$  is locally nilpotent.

*Theorem 35. (Kostrikin)* A Lie ring  $L$  of characteristic  $p$  satisfying the  $m$ th Engel condition,  $m < p$ , is locally nilpotent. Consequently the restricted Burnside problem is answered affirmatively for prime exponent  $p$ .

For exponent  $p^s$ ,  $s > 1$ , present information on the restricted Burnside problem is very unsatisfactory. The commutator relations known do not yield a genuine Engel condition. The additive order of elements in the associated Lie ring is  $p^s$  rather than  $p$ , and this introduces complications.

In groups of exponent 4, since Sanov has proved the finiteness of the groups  $B(4, k)$ , the restricted and unrestricted Burnside problems coincide. C. R. B. Wright [50] has shown that with  $G = B(4, k)$  the class is at most  $3k - 1$ , showing that  $G_{3k} = G_{3k+1} = \dots$ , whence by Sanov's result  $G_{3k} = 1$ . This of course provides a far more satisfactory basis for estimating the order of  $B(4, k)$  than Sanov's original work. For exponent 4 we find

$$(8.21) \quad (y, x, x, x, x, x) = 1$$

for any  $x$  and  $y$ . This is the fifth Engel condition. The writer, in work as yet unpublished, has found a kind of third Engel condition, specifically

$$(8.22) \quad (x, y, z, w, w, w) = 1 \pmod{G_7}$$

Wright's main result is that

$$(8.23) \quad (a_1, a_2, \dots, a_n) \equiv 1 \pmod{G_{n+1}}$$

providing, (1)  $n \geq 6$  and (2) four or more of the  $a_i$  are equal. This easily gives the result that the class of  $B(4, k)$  is at most  $3k$ , and a special argument reduces this to  $3k - 1$ .

*Problem 10. Develop a more precise relationship between Lie rings and nilpotent groups. Different groups may correspond to the same Lie ring.*

*Problem 11. Using (8.6) rather than (8.7), develop a Lie theory for the Burnside problem rather than merely the restricted Burnside problem.*

*Problem 12. Investigate fully invariant subgroups of free groups. These are all "word groups." In particular they include Bruck's [6] groups of type  $(g \rightarrow c)$ , those groups in which every commutator of weight  $c + 1$  in  $g$  elements is the identity.*

## REFERENCES

1. R. Baer and F. Levi, "Freie Produkte und ihre Untergruppen," *Comp. Math.* **3** (1938) 391-398.
2. G. Baumslag, W. W. Boone, and B. H. Neumann, "Some unsolvable problems about elements and subgroups of groups," *Math. Scand.* **1** (1959) 191-201.
3. W. W. Boone, "The word problem," *Ann. of Math.* **70** (1959) 207-265.
4. G. Birkhoff, "Representability of Lie algebras and Lie group by matrices," *Ann. of Math.* (2) **38** (1937) 526-532.
5. J. H. Britton, "The word problem for groups," *Proc. London Math. Soc.* (3) **8** (1948) 493-506.
6. R. H. Bruck, *What makes a group finite?* To be published by Prentice-Hall, Englewood Cliffs, N.J.
7. W. Burnside, "On an unsettled question in the theory of discontinuous groups," *Quart. J. Pure and Appl. Math.* **33** (1902) 230-238.
8. H. S. M. Coxeter and W. O. J. Moser, *Generators and relations for discrete groups*, Berlin (1957) Springer.
9. R. Dean, "A sequence without repeats on  $x, x^{-1}, y, y^{-1}$ ," to appear in *Amer. Math. Monthly*.
10. F. Gross, *On the 2-length of solvable groups*, Doctoral Thesis, California Institute of Technology (1964).
11. M. Hall, Jr., "A topology for free groups and related groups," *Ann. of Math.* **52** (1950) 127-139.
12. ———, "A basis for free Lie rings and higher commutators in free groups," *Proc. Amer. Math. Soc.* **1** (1950) 575-581.
13. ———, "Subgroups of free products," *Pacific J. Math.* **3** (1953) 115-120.
14. ———, "Solution of the Burnside problem for exponent 6," *Illinois J. Math.* **2** (1958) 764-786.

15. ———, "Coset Representatives in free groups," *Trans. Amer. Math. Soc.* **67** (1949) 421-432.
16. ———, *The Theory of Groups*, New York (1959) Macmillan.
17. M. Hall, Jr. and T. Rado, "On Schreier systems in free groups," *Trans. Amer. Math. Soc.* **64** (1948) 386-408.
18. P. Hall, "A contribution to the theory of groups of prime-power order," *Proc. London Math. Soc.* **36** (1933) 29-95.
19. P. Hall and G. Higman, "The  $p$ -length of a  $p$ -soluble group, and reduction theorems for Burnside's problem," *Proc. London Math. Soc.* (3) **7** (1956) 1-42.
20. H. Heineken, "Engelsche Elemente der Lange drei," *Illinois J. Math.* **5** (1961) 681-707.
21. P. Higgins and R. Lyndon, "Equivalence of elements under automorphisms of a free group," to be published.
22. G. Higman, "On finite groups of exponent five," *Proc. Cambridge Philos. Soc.* **52** (1956) 381-390.
23. ———, "Lie ring methods in the theory of finite nilpotent groups," *Proc. Intern. Congr. Math.* 1958.
24. ———, "Subgroups of finitely presented groups," *Proc. Roy. Soc. London Ser. A* **262** (1961) 455-475.
25. A. H. M. Hoare, "A note on 2-soluble groups," *J. London Math. Soc.* **35** (1960) 193-199.
26. A. I. Kostrikin, "Solution of the restricted Burnside problem for the exponent 5," *Izv. Akad. Nauk SSSR* **19** (1955) 233-244 (Russian).
27. ———, "On locally nilpotent Lie rings satisfying an Engel condition," *Dokl. Akad. Nauk SSSR* **118** (1958) 1074-1077 (Russian).
28. ———, "On Burnside's problem," *Dokl. Akad. Nauk SSSR* **119** (1958) 1081-1084.
29. M. Lazard, "Sur les groupes nilpotents et les anneaux de Lie," *Ann. Sci. Ec. Norm. Sup.* (3) **71** (1954) 101-190.
30. F. W. Levi and B. L. van der Waerden, "Uberline besondere Klasse von Gruppen," *Math. Zeit.* **32** (1930) 315-318.
31. S. MacLane, "A proof of the subgroup theorem for free products," *Mathematika* **5** (1958) 13-19.
32. W. Magnus, "Die Idenitatsproblem fur Gruppen mit einer definierenden Relation," *Math. Ann.* **106** (1932) 295-307.
33. M. Morse and G. A. Hedlund, "Unending chess, symbolic dynamics and a problem in semigroups," *Duke Math. J.* **11** (1944) 1-7.
34. B. H. Neumann, "An Essay on free products of groups with amalgamations," *Philos. Trans. Roy. Soc. London Ser. A* **246** 503-554.

# 3

## Some Aspects of the Topology of 3-Manifolds Related to the Poincaré Conjecture

*R. H. Bing*

### 1. INTRODUCTION

One of the most interesting questions in the study of 3-manifolds asks if the following conjecture is true.

*Poincaré Conjecture.* A compact connected 3-manifold is topologically a 3-sphere if it is simple connected.

Our discussion will center around some of the aspects of 3-manifolds as they relate to this conjecture.

The last decade has seen the solutions of many interesting and important problems about 3-manifolds. When we consider some of the remarkable results obtained, it is a bit surprising that the Poincaré conjecture remains unconquered. Before reviewing some definitions at the end of this section, we mention several of these problems.

Moise showed [15] that for open subsets  $U$  of Euclidean 3-space  $E^3$ , homeomorphisms can be approximated by piecewise linear homeomorphisms. If  $U$  is an open subset of  $E^3$ ,  $f$  is a continuous positive function defined on  $U$ , and  $h: U \rightarrow E^3$  is a homeomorphism, there is a piecewise linear homeomorphism  $g: U \rightarrow E^3$  such that

distance  $[h(x), g(x)] < f(x)$  for each point  $x$  of  $U$  (see Figure 1). This result was used to show [16] that 3-manifolds can be triangulated. The result also yields the Hauptvermutung. If  $T_1, T_2$  are two triangulations of the same 3-manifold, then  $T_1, T_2$  have isomorphic rectilinear subdivisions. Hence, in attacking the Poincaré conjecture, it suffices to limit our consideration to piecewise linear 3-manifolds.

Papakyriakopoulos [18] made a remarkable breakthrough when he proved Dehn's lemma, which states that if  $D$  is a 2-simplex,  $M^3$  is a combinatorial 3-manifold, and  $f: D \rightarrow M^3$  is a piecewise linear map such that for some neighborhood  $N$  of  $Bd D$  in  $D$ ,  $f^{-1}: f(N) \rightarrow D$  is a homeomorphism, then  $f(Bd D)$  bounds a polyhedral disk in  $M^3$  (see Figure 2). The theorem was proved by scissors and paste techniques. Using similar techniques, Papa established [17] the loop theorem—if  $M^3$  is a triangulated 3-manifold with boundary and  $L$  is a polyhedral singular closed curve on  $Bd M^3$  that can be shrunk to a point in  $M^3$  but not on  $Bd M^3$ , then there is a polyhedral simple closed curve with the same property.

Another noteworthy theorem by Papa [18] was the sphere theorem, which states that if  $S^2$  is a 2-sphere,  $M^3$  a compact triangulated 3-manifold, and  $f: S^2 \rightarrow M^3$  a map of  $S^2$  into  $M^3$  that is not homotopic to a constant, there is a polyhedral 2-sphere in  $M^3$

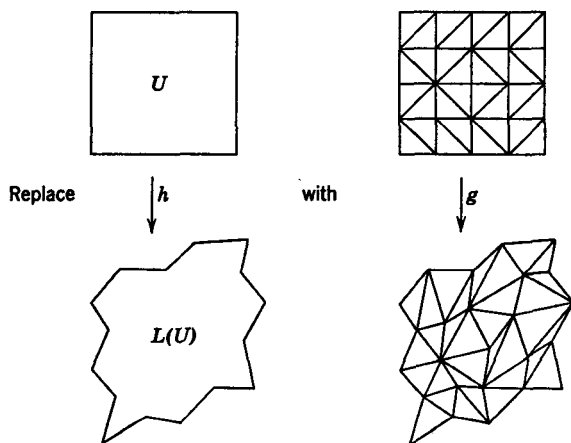


Figure 1

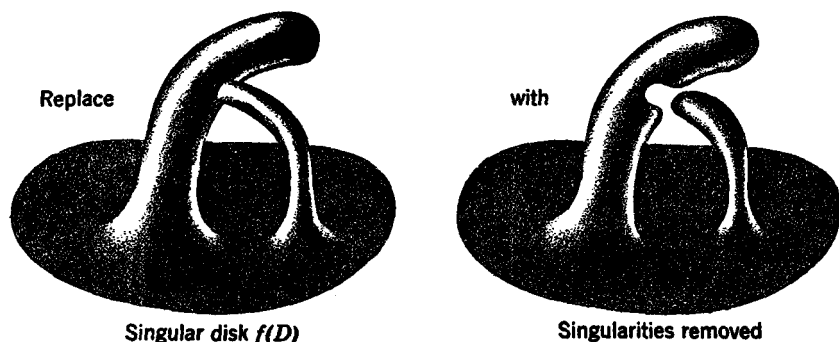


Figure 2

which cannot be shrunk to a point in  $M^3$ . We might hope that this Sphere theorem would be of some help in classifying compact 3-manifolds, but again the Poincaré conjecture gets in the way [19, 14]. Papa made an attack [20] on the Poincaré conjecture, but the victory was only partial.

Bing partially bridged the gap between the study of the topological embeddings of objects in triangulated 3-manifolds and the piecewise linear embeddings of these objects when he showed [4] that any surface in  $E^3$  can be approximated by a polyhedral surface. If  $M^2$  is a 2-manifold,  $f$  is a continuous nonnegative function defined on  $M^2$ , and  $h: M^2 \rightarrow E^3$  is a homeomorphism of  $M^2$  onto a closed subset of  $E^3$ , there is a piecewise linear homeomorphism  $g: M^2 \rightarrow E^3$  such that for each point  $x \in M^2$ , distance  $[h(x), g(x)] < f(x)$  and  $g$  is locally piecewise linear at  $x \in M^2$  if  $f(x) > 0$ . Figure 3 indicates how to start building an approximation to  $h(M^2)$ . This theorem (with later refinements of it) were used to give [6] an alternative proof of Moise's results on the triangulation of 3-manifolds and of his results that homeomorphisms of open subsets of  $E^3$  can be approximated by piecewise linear homeomorphisms.

Bing made an attack on the Poincaré conjecture in [5] but achieved only a partial solution. Other partial solutions not previously appearing in the literature are provided by Theorems 7, 17, and 18 of this chapter.

An  $n$ -manifold is a separable metric space, each of whose points has a neighborhood homeomorphic with Euclidean  $n$ -space  $E^n$ . Some mathematicians denote  $E^n$  by  $R^n$  since it is the cartesian



together with some particular triangulation of the geometric complex. To triangulate a manifold, we may need to re-metrize it so that we have the sum of simplexes of the right sort. When we speak of a triangulation of  $E^n$ , however, unless we indicate to the contrary, we mean a rectilinear triangulation of  $E^n$  where the simplexes of the triangulation are simplexes under the ordinary Euclidean metric.

Suppose  $T_1$  is a triangulation of a geometric complex  $C$ . A sub-division  $T_2$  of  $T_1$  is a triangulation of  $C$  (keeping the same metric for  $C$ ) such that each simplex of  $T_2$  lies in a simplex of  $T_1$ .

The  $i$ -skeleton of a triangulation is the sum of all  $i$ -simplexes in it.

A geometric complex may have several triangulations. A set  $Y$  in a geometric complex is called a *polyhedron* if the geometric complex has a triangulation (without re-metrization of the complex) such that  $Y$  is the sum of simplexes of the triangulation. A set  $Z$  is *tame* in a geometric complex  $C$  if there is a homeomorphism  $h$  of  $C$  onto itself such that  $h(Z)$  is a polyhedron. A closed set  $X$  in  $C$  which is not tame but is homeomorphic with a polyhedron is called *wild*. A set  $X$  is *locally tame* at a point  $p$  of an  $n$ -manifold if there is a neighborhood  $N$  of  $p$  and a homeomorphism  $h: \tilde{N} \rightarrow E^n$  such that  $h(\tilde{N} \cdot X)$  is a polyhedron in  $E^n$ .

A map  $f$  of a geometric complex  $C$  into a geometric complex  $C'$  is *piecewise linear* if there is a triangulation  $T$  of  $C$  such that  $f$  is linear on each simplex of  $T$ . The map is called *locally piecewise linear* at a point  $p$  if for some neighborhood  $N$  of  $p$  there is a triangulation  $T'$  of  $\tilde{N}$  such that  $f$  is linear on each element of  $T'$ . If  $f$  is locally piecewise linear except possibly at points of a set  $X$ , we say that it is *locally piecewise linear mod  $X$* .

A triangulated  $n$ -manifold  $M^n$  is a *piecewise linear manifold* if for each point  $p$  of  $M^n$  there is a piecewise linear homeomorphism of  $E^n$  onto a neighborhood of  $p$  in  $M^n$ . Although it is known that any triangulation of a 3-manifold yields a piecewise linear manifold, the corresponding result is not known for triangulated 4-manifolds (see Section 8).

If  $s$  is a simplex of a triangulation  $T$  of a geometric complex, *Star* ( $s, T$ ) is the sum of all simplexes of  $T$  that contain  $s$ . If  $s, s'$  are mutually exclusive simplexes that are faces of the same simplex  $s''$ , the *join* of  $s$  and  $s'$  is the sum of all segments in  $s''$  from a point of  $s$  to a point of  $s'$ . *Link* ( $s, T$ ) is the sum of all simplexes  $s'$  of  $T$  such that the join of  $s$  and  $s'$  is a simplex of  $T$ . Some authors refer to

the link of  $s$  as the *complement* of  $s$ . When  $T$  is a triangulation of a manifold and  $s$  is a vertex of  $T$ ,  $\text{Link}(s, T) = \text{Bd Star}(s, T)$ .

We use  $I^n$  to denote an  $n$ -cell. A topological set  $X$  is  *$n$ -connected* if each map of  $\text{Bd } I^{n+1}$  into  $X$  can be extended to map  $I^{n+1}$  into  $X$ . This is a bit more delicate than terminology used by some authors who say that a set is  *$i$ -connected* if and only if for each  $j \leq i$ , the set is  *$j$ -connected* as we have defined it. A set is called *simply connected* if it is 1-connected. A set  $U$  is *locally  $n$ -connected* at a point  $p$  of  $\bar{U}$  if for each neighborhood  $N_1$  of  $p$  there is a neighborhood  $N_2$  of  $p$  such that each map of  $\text{Bd } I^{n+1}$  into  $N_2 \cdot U$  can be extended to map  $I^{n+1}$  into  $N_1 \cdot U$ . Note that  $p$  need not be a point of  $U$ .

A compact connected simply connected 3-manifold is called a *homotopy 3-sphere*. The Poincaré conjecture claims that a homotopy 3-sphere is a 3-sphere, but the issue is still in doubt.

A *homotopy* is a continuous 1-parameter family of maps  $f_t (0 \leq t \leq 1)$ . If  $f: X \times [0, 1] \rightarrow Y$ , we find it convenient to denote  $f(x \times t)$  by  $f_t(x)$ . We say that the maps  $g_1: X \rightarrow Y$ ,  $g_2: X \rightarrow Y$  are *homotopic in  $Y$*  provided there is a map  $f: X \times [0, 1] \rightarrow Y$  such that  $f_0 = g_1$ ,  $f_1 = g_2$ . The homotopy is called an *isotopy* if each  $f_t$  is a homeomorphism.

A constant map of  $X$  is a map that sends all of  $X$  into a single point. We say that a set  $X$  can be shrunk to a point in a set  $Y$  if there is a map  $f: X \times [0, 1] \rightarrow Y$  such that  $f_0 = I$  (identity map) and  $f_1$  is a constant map.

## 2. EXAMPLES OF COMPACT 3-MANIFOLDS

Perhaps the best known example of a 3-manifold is Euclidean 3-space  $E^3$  itself. Space  $E^3$  is not compact, however, since it has an infinite set of points with no limit point. Examples of compact 3-manifolds are somewhat artificial since there are no physical objects in  $E^3$ , the world in which we presume we live, which are examples of compact 3-manifolds. We do have objects in  $E^3$  which are examples of compact 3-manifolds with boundaries such as a cube, a solid torus, a cube with handles, a cube with a knotted hole, and a punctured cube. We also have examples of compact 2-manifolds. In this section we use these real examples to describe examples of compact 3-manifolds which are abstract in the sense that in  $E^3$ , there are no physical objects topologically equivalent to them.

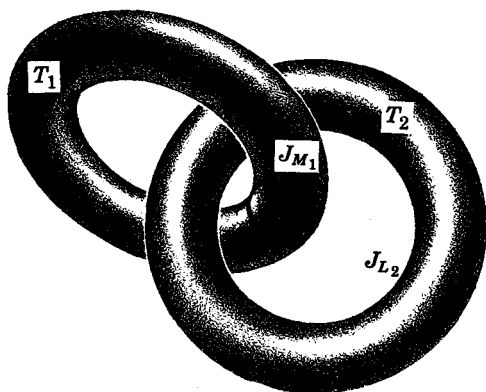
Sewing 3-manifolds with boundaries together is a convenient way of constructing examples of 3-manifolds. Suppose  $X, Y$  are two topological sets and  $h$  is a homeomorphism of a closed subset  $A$  of  $X$  onto a closed subset  $h(A)$  of  $Y$ . Then  $X +_h Y$  denotes the topological set obtained by sewing  $X$  and  $Y$  together along  $A$  and  $h(A)$  so that a point of  $A$  is identified with its image under  $h$ . It is convenient to think of  $X$  and  $Y$  as intersecting only in  $A = h(A)$  so that  $a$  coincides with  $h(a)$ . An open set in  $X +_h Y$  is either an open set in  $X - A$ , an open set in  $Y - h(A)$ , or a set of the form  $U_1 + U_2$  where  $U_1$  is an open set in  $X$ ,  $U_2$  is an open set in  $Y$ , and  $h(U_1 \cdot A) = U_2 \cdot h(A)$ .

1.  $S^n$ . An  $n$ -sphere  $S^n$  is a set homeomorphic to the boundary of a unit  $(n + 1)$ -ball in  $E^{n+1}$ . It might have the equation  $x_1^2 + x_2^2 + \cdots + x_{n+1}^2 = 1$ . Similarly,  $S^n$  is a set homeomorphic with  $Bd I^{n+1}$ .

2.  $S^3$ . Since  $E^4$  does not seem as real to many of us as  $E^3$ , it may be more convenient to regard  $S^3$  as the one-point-compactification of  $E^3$ . Here, a deleted neighborhood of the ideal point at infinity is the exterior of a closed 3-ball—the larger the 3-ball, the smaller the neighborhood.

3.  $S^3$  as a toroidal manifold. Suppose  $T$  is the solid torus obtained by revolving in  $E^3$  the disk  $(x - 2)^2 + y^2 \leq 1$  in the  $xy$ -plane about the  $y$ -axis. The simple closed curve  $J_M$  defined by  $z = 0, (x - 2)^2 + y^2 = 1$  bounds a disk in  $T$  but not a disk on  $Bd T$  and is called a meridional simple closed curve of  $Bd T$ . The simple closed curve  $J_L$  defined by  $y = 0, x^2 + z^2 = 1$  bounds a 2-manifold in  $E^3 - \text{Int } T$  but none on  $Bd T$  and is called a longitudinal simple closed curve of  $Bd T$ . Suppose  $T_1, T_2$  are two copies of  $T$  with  $J_{M1}, J_{L1}, J_{M2}, J_{L2}$  being the corresponding meridional and longitudinal simple closed curves. If  $h$  is a homeomorphism of  $Bd T_1$  onto  $Bd T_2$  that takes  $J_{M1}$  onto  $J_{L2}$ , then  $T_1 +_h T_2$  is topologically  $S^3$ . It may be convenient to regard  $T_1, T_2$  as being two linked inner tubes plus their interiors (see Figure 4). The point at infinity is moved to be on the interior of  $T_2$ . Then the two inner tubes are blown up so that  $Bd T_1$  and  $Bd T_2$  coincide and  $T_1 + T_2$  fills up  $E^3$  plus the point added at infinity.

4. Toroidal manifold  $T(m, n)$ . Suppose that  $m, n$  are relatively prime positive integers;  $T_1, T_2$  are as in the preceding example; and  $p_1, p_2, \dots, p_n$  are  $n$  points on  $J_{M2}$  cyclicly ordered on  $J_{M2}$  as



$$T_1 +_h T_2 = S^3$$

Figure 4

indicated by their subscripts. Let  $J(m, n)$  be a simple closed curve on  $Bd T_2$  that intersects  $J_{M_2}$  only at  $p_1, p_2, \dots, p_n$ ; crosses  $J_{M_2}$  at each such point of intersection, with the  $p_i$ 's cyclicly ordered on  $J(m, n)$  as  $p_1, p_{m+1}, p_{2m+1}, \dots, p_{(n-1)m+1}$  (see Figure 5). Let  $h$  be a homeomorphism of  $Bd T_1$  onto  $Bd T_2$  taking  $J_{M_1}$  onto  $J(m, n)$ . Then  $T_1 +_h T_2 = T(m, n)$ . It is known that  $T(1, 1)$  is the only simply connected toroidal manifold. As pointed out in the previous example,  $T(1, 1)$  is topologically  $S^3$ . Hence toroidal manifolds do not furnish a clue as to whether or not the Poincaré conjecture is true. Had  $T(0, 1)$  been defined, it would represent the cartesian product  $S^2 \times S^1$ .

5. Double toroidal manifolds. Suppose that in  $E^3$  that  $X$  is a tubular neighborhood of a figure eight. It may be regarded as a cube with two handles. Suppose  $X_1, X_2$  are two copies of  $X$  and  $h$  is a homeomorphism of  $Bd X_1$  onto  $Bd X_2$ . Then  $X_1 +_h X_2$  is a double toroidal manifold. It is not known whether or not each simply connected double toroidal manifold is topologically  $S^3$ .

6.  $n$ -tuple toroidal manifolds. Suppose  $X$  be a cube with  $n$  handles and  $X_1, X_2$  are two copies of  $X$ . For each homeomorphism  $h: Bd X_1 \rightarrow Bd X_2$ ,  $X_1 +_h X_2$  is an  $n$ -tuple toroidal manifold. It is known that each compact connected orientable (does not contain a solid Kline bottle) 3-manifold can be regarded as an  $n$ -tuple toroidal

manifold. To regard it as such, consider a triangulation of the manifold and let  $X_1$  be a tubular neighborhood of its 1-skeleton. If there is a counterexample to the Poincaré conjecture, it lies among the  $n$ -tuple toroidal manifolds. It would be interesting to know for what integers  $n$  there is no counterexample. This attack seems difficult since we have only found the answer for  $n = 1$ .

7. Homology spheres. Let  $C_1, C_2$  be cubes with knotted holes (see Figure 6). Let  $J_{M_i}$  be a simple closed curve on  $Bd C_i$  such that  $J_{M_i}$  bounds an orientable 2-manifold in  $C_i$  but not on  $Bd C_i$  and  $J_{L_i}$  be another simple closed curve on  $Bd C_i$  that intersects  $J_{M_i}$  at just one point and crosses it there. If  $h: Bd C_1 \rightarrow Bd C_2$  is a homeomorphism that takes  $J_{M_1}, J_{L_1}$  onto  $J_{L_2}, J_{M_2}$  respectively, then  $C_1 +_h C_2$  is a 3-manifold with the same homology as  $S^3$ , but which is not a homotopy sphere. The reason that  $C_1 +_h C_2$  is not a homotopy sphere is as follows. If a tame 2-dimension torus separates a homotopy 3-sphere, it follows from the Loop theorem and Dehn's Lemma that one of the pieces into which the torus separates the homotopy 3-sphere has the same fundamental group as a circle. These homology spheres we have described give counterexamples to the Poincaré conjecture in its original form (see Section 8).

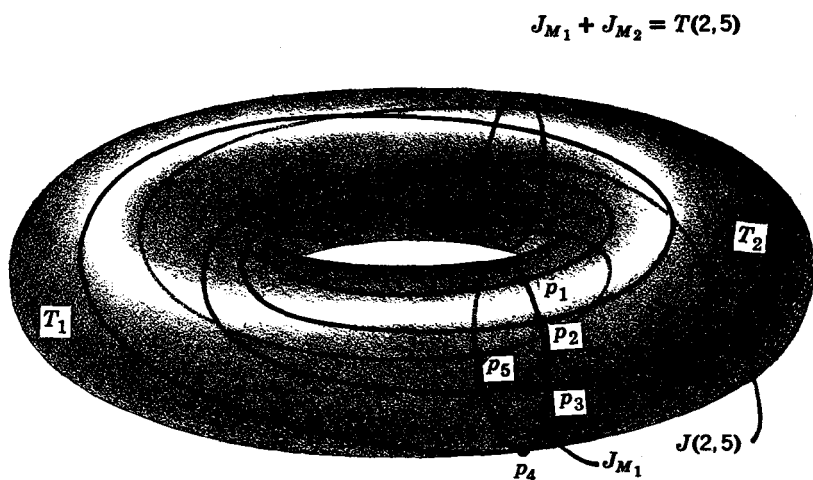


Figure 5

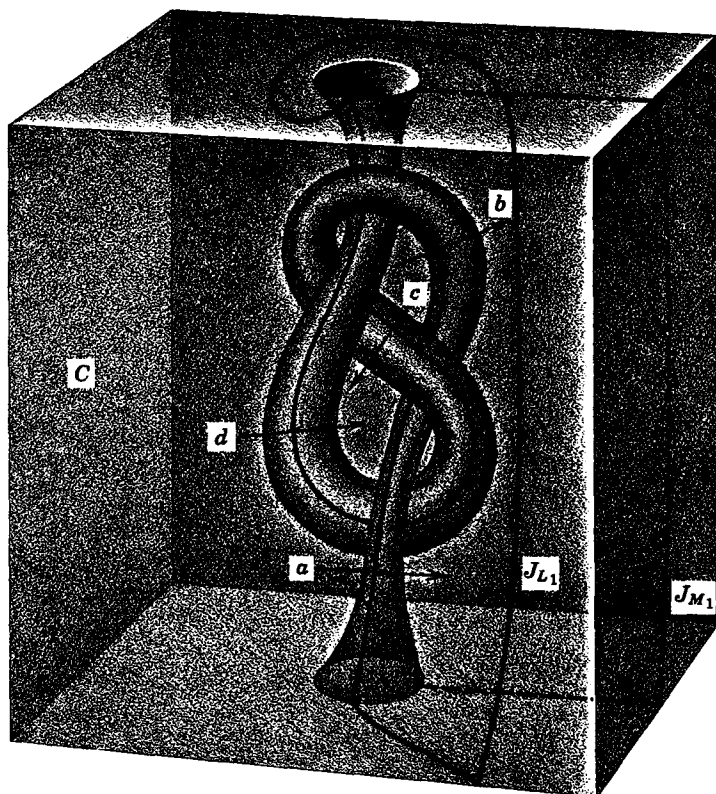


Figure 6

8. A candidate for a counterexample. Suppose  $J_{L_1}$  is selected in the  $Bd C_1$  of Figure 6 so that  $J_{L_1}$  goes through the hole once and does not bound in  $E^3 - \text{Int } C_1$ . Let  $T$  be a solid torus and  $h: Bd C_1 \rightarrow Bd T$  be a homeomorphism that takes  $J_{L_1}$  onto a meridional simple closed curve on  $Bd T$ . Is  $C_1 +_h T$  a counterexample to the Poincaré conjecture? The fundamental group of  $C_1$  can be found to be  $G = \{a, b, c, d / cac^{-1}b^{-1} = bdc^{-1}d^{-1} = aca^{-1}d^{-1} = 1\}$ . After  $T$  is sewn in, the fundamental group of the sum is  $G' = \{cac^{-1}b^{-1} = bdc^{-1}d^{-1} = aca^{-1}d^{-1} = c^{-1}da^{-1}ba = 1\}$ . Whether or not  $C_1 +_h T$  is a homotopy sphere depends on whether or not  $G'$  is trivial.

Perhaps by drilling knotted holes (possibly different from the one

shown in Figure 6) and sewing back in solid tori, we obtain a group whose relations, through some quirk, makes the group trivial but obtains a sum that is not  $S^3$ . Two difficulties serve as blocks to this approach. First, the algebra is fierce and it seems only by chance that we can show a group to be trivial. Second, even if the algebra succeeds, we still need to show that the resulting space is topologically different from  $S^3$ . Bing has shown [5] that no counterexample results if the knot in  $C_1$  is a trefoil knot, and Hempel [11] has extended this result by showing that no counterexample results for certain other knots. The four-crossing knot shown in Figure 6 still baffles us.

9. Projective 3-space  $P^3$ . Suppose  $B$  is a unit ball in  $E^3$  and  $P^3$  is the space obtained by sewing each point  $p$  of  $B$  onto its diametrically opposite point (antipodal point). Note that under the identification,  $B$  goes onto a projective plane which is a one sided surface in  $P^3$ . It follows readily that  $P^3$  is not simply connected since this projective plane does not separate  $P^3$ . In a simply connected 3-manifold, each 2-manifold separates it.

10. Lens space  $L(m, n)$ . Let  $B$  be the unit ball in  $E^3$  described by the equation  $x^2 + y^2 + z^2 \leq 1$ ;  $m, n$  be relatively prime positive integers; and  $L(m, n)$  be the decomposition space of  $B$  obtained by identifying the upper half of  $B$  with the lower half under a rotation through  $2m\pi/n$  radians and a reflection through the  $z = 0$  plane. Each point on  $z = 0$ ,  $x^2 + y^2 = 1$  is identified with  $n - 1$  other points; each point of  $x^2 + y^2 + z^2 = 1$ ,  $z > 0$  is identified with one of  $x^2 + y^2 + z^2 = 1$ ,  $z < 0$ ; and points of  $x^2 + y^2 + z^2 < 1$  are left alone. It may be shown that each lens space is a toroidal manifold. The tori correspond to tubular neighborhoods of the axis and the equator. The upper and lower halves of  $B$  shown in Figure 7 are shown making a  $72^\circ$  angle with each other so that when sewing is done the tubular neighborhood of the equator looks like a torus without unnecessary distortion. Note that  $L(1, 1)$  is topologically equivalent to  $S^3$ . In fact  $L(m, n)$  is topologically equivalent to  $T(m, n)$ . Each lens space is a toroidal manifold and  $S^2 \times S^1$  is the only toroidal manifold which is not a lens space.

11. Cartesian products. Taking the cartesian product of  $S^1$  with various compact 2-manifolds provides us with an abundant supply of compact 3-manifolds. None of these are homotopy spheres.

12. Replacing tori. It is possible to obtain a set topologically equivalent to  $S^2 \times S^1$  by removing an unknotted solid torus from

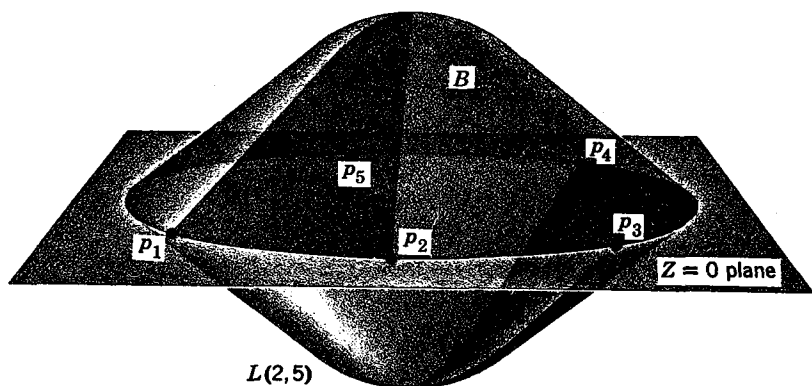


Figure 7

$S^3$  and sewing it back in so that the meridional and longitudinal simple closed curves on the boundary are interchanged. In fact, A. H. Wallace [28] and W. B. R. Lickorish [10] have shown that any compact connected orientable 3-manifold whatever can be obtained by removing a finite collection of mutually exclusive solid tori and sewing them in differently.

13. R. H. Fox has suggested [8] branched covering spaces of  $S^3$  as a possible place to look for counterexamples to the Poincaré conjecture.

### 3. IMPROVING TRIANGULATIONS

Consider a 2-dimensional torus with a small hole. If the hole is large, we can easily see how to triangulate the 2-manifold with boundary so that all the vertices lie on the boundary. If the hole is small, the result remains true but is more surprising.

There are many different ways of triangulating the same 3-manifold with boundary. If we triangulate a simply connected 3-manifold and remove the interior of one 3-simplex, what is left is called a *fake cube*. Like a 3-cell, its boundary is a 2-sphere and it is connected as well as simply connected. If it is a 3-cell, the original 3-manifold was a 3-sphere, but if it is not a 3-cell, the original 3-manifold was a counterexample to the Poincaré conjecture.

It might conceivably be of some help in deciding whether a fake

cube is a real cube if the fake cube has a "nice" triangulation. Theorem 1 shows that each fake cube has a triangulation all of whose vertices lies on the boundary. The theorem can be proved for arbitrary compact, connected, combinatorial  $n$ -manifolds, but for simplicity we only discuss the case  $n = 3$ .

*Theorem 1.* Each compact connected 3-manifold  $M^3$  with non-null boundary has a triangulation all of whose vertices are on  $Bd M^3$ .

PROOF. It follows from [3] that  $M^3$  has a triangulation  $T'$ . If a vertex of  $T'$  lies in  $\text{Int } M^3$ , we use two steps to reduce the number of vertices in  $\text{Int } M^3$ . Let  $vv'$  be an edge of  $T'$  with  $v \in Bd M^3$  and  $v' \in \text{Int } M^3$ . If  $\text{Link}(v', T')$  has a 2-simplex on  $Bd M^3$ , we can skip the first step.

1. Let  $U$  be the open half-star of  $v$ —that is, the set of all points of  $\text{Star}(v, T')$  which lie closer to  $v$  than to the other end of the segment through them from  $v$  to  $\text{Link}(v, T')$ . Note that  $M^3 - U$  is topologically  $M^3$ . However,  $T'$  is not a triangulation of  $M^3 - U$  since some of the original 3-simplexes of  $T'$  intersect  $M^3 - U$  in a tetrahedron with one end blunted. Suppose  $T''$  is a triangulation of  $M^3 - U$  obtained by leaving the elements of  $T'$  in  $M^3 - \text{Star}(v, T')$  alone and subdividing the blunted tetrahedra so as not to introduce new corners and so that some 3-simplex of  $T''$  with a face on  $Bd(M^3 - U)$  has what is left of  $vv'$  as an edge (see Figure 8).

2. Let  $s^3$  be a 3-simplex of  $T''$  with  $v'$  as vertex and with a face on  $Bd(M^3 - U)$ . Denote the closure of  $(M^3 - U) - s^3$  by  $N^3$ . Then  $T''$  has one fewer vertex in  $\text{Int } N^3$  than  $T'$  has in  $\text{Int } M^3$ . Since  $N^3$  is topologically equivalent to  $M^3$ , there is a triangulation of  $M^3$  that has fewer vertices on  $\text{Int } M^3$  than  $T'$  does. An iteration of this procedure yields  $T$ .

*Theorem 2.* Suppose  $T$  is a triangulation of a compact 3-manifold with boundary  $M$  such that each vertex of  $T$  lies on  $Bd M$  and each 3-simplex of  $M$  has at least two edges on  $Bd M$ . Then  $M$  has a triangulation  $T'$  such that each vertex of  $T'$  lies on  $Bd M$  and each 2-simplex of  $T'$  has an odd number of edges on  $Bd M$ .

PROOF. Suppose  $T$  is chosen so as to have a minimal number of 2-simplexes not on  $Bd M$  while maintaining the property that all vertices of  $T$  lie on  $Bd M$  and each 3-simplex of  $T$  has at least two edges on  $Bd M$ . We show that this minimal triangulation of  $M$  suffices for  $T'$ .

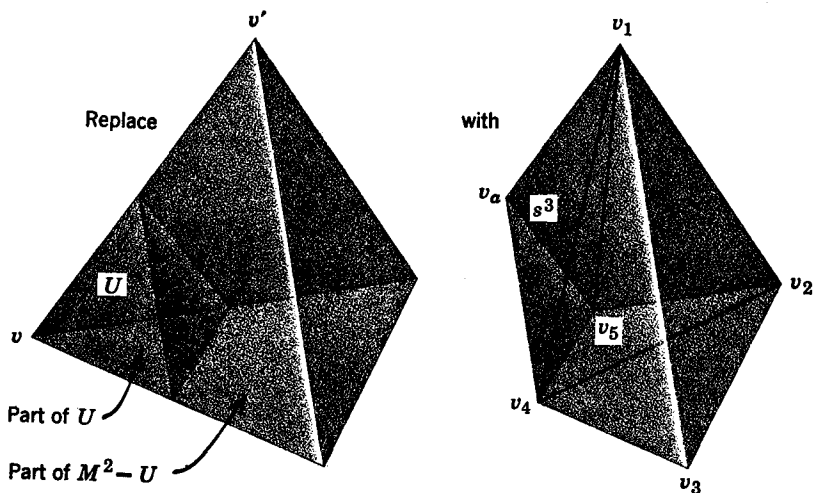


Figure 8

Assume that  $T$  has a 2-simplex  $v_1v_2v_3$  with precisely two edges  $v_1v_2, v_1v_3$  on  $Bd M$ . We split  $M$  along this 2-simplex (see Figure 9). Let  $X$  be the star of  $v_1$  with respect to  $T$  (sum of all simplexes of  $T$  that contain  $v_1$ ). Then  $X$  is a cube which is the sum of two cubes  $X_1, X_2$  such that  $X_1 \cdot X_2 = v_1v_2v_3$ . Split  $M$  along  $v_1v_2v_3$  to get a 3-manifold with boundary homeomorphic with  $M$  with a triangula-

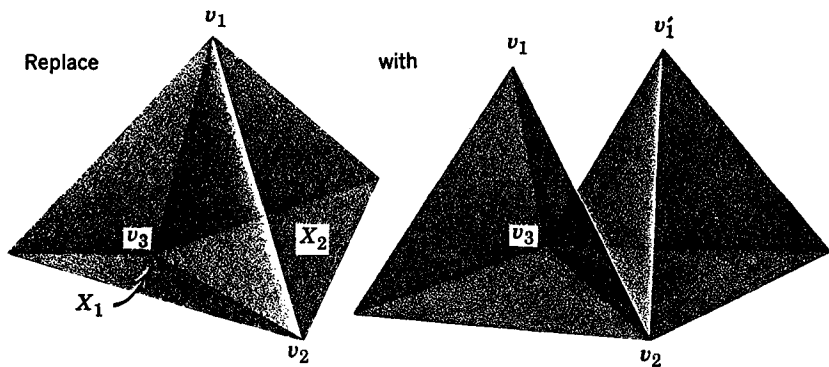


Figure 9

tion  $T_1$  such that a simplex of  $T_1$  is either a simplex of  $T$  that misses  $v_1$ , a simplex of  $T$  in  $X_1$ , or a simplex of  $T$  in  $X_2$  with the vertex  $v_1$  replaced with a new vertex  $v'_1$ . Note that  $T'$  has an extra vertex  $v'_1$ , two extra edges  $v'_1v_2$ ,  $v'_1v_3$ , and an extra face  $v'_1v_2v_3$ , but it has one fewer 2-simplex not on  $Bd M$ . Since  $T$  was already minimal in this respect, there is no  $v_1v_2v_3$  as assumed at the beginning of this paragraph.

We now show that under the triangulation  $T$  we have chosen, each 2-simplex of  $T$  has at least one edge on  $Bd M$ . Suppose  $v_1v_2v_3$  is a 2-simplex of  $T$  with no edge on  $T$ ; then if  $v_1v_2v_3v_4$  is a 3-simplex of  $T$ ,  $v_1v_2v_3v_4$  has at least two edges on  $Bd M$  by hypothesis and one of the three 2-simplexes  $v_2v_3v_4$ ,  $v_1v_3v_4$ ,  $v_1v_2v_4$  has precisely two edges on  $Bd M$ . This was ruled out in the preceding paragraph.

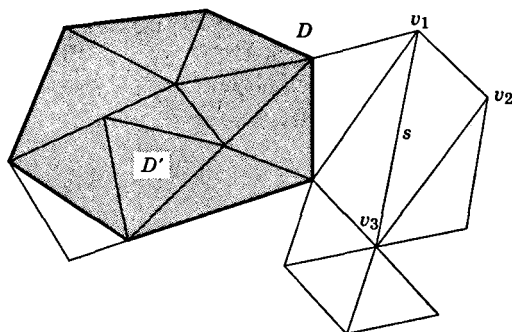
#### 4. A FAKE CUBE IS REAL IF IT HAS A "NICE" TRIANGULATION

A natural way of attempting to prove the Poincaré conjecture is to try to show that a fake cube has a triangulation that can be shelled. We say that a triangulation  $T$  of an  $n$ -cell can be *shelled* if the  $n$ -simplexes of  $T$  can be ordered  $s_1, s_2, \dots, s_n$  so that for each integer  $k \leq n$ ,  $s_k + s_{k+1} + \dots + s_n$  is an  $n$ -cell. A triangulation of a fake cube can be *shelled* if the 3-simplexes of the triangulation can be ordered  $s_1, s_2, \dots, s_n$ , such that for each  $k \leq n$ ,  $s_k + s_{k+1} + \dots + s_n$  is a fake cube.

*Theorem 3.* Suppose  $T$  is a triangulation of a 2-cell  $D$  and  $D'$  is a disk in  $D$  which is the sum of elements of  $T$ . Then the 2-simplexes of  $T$  not contained in  $D'$  can be ordered  $s_1, s_2, \dots, s_n$  so that for each integer  $k \leq n$ ,  $D' + s_k + s_{k+1} + \dots + s_n$  is a disk.

*PROOF.* The theorem is proved by induction on  $n$ . Let  $s$  be a 2-simplex of  $T$  which does not lie in  $D'$  but which has an edge on  $Bd D$ . We can either let  $s$  be  $s_1$  and shell it or else some edge  $s'$  of  $s$  not on  $D'$  is a spanning arc of  $D$ . For the latter we consider the two disks into which  $s'$  separates  $D$  and expand  $D'$  to be one of them. We have thereby reduced  $n$  and our inductive process is started (see Figure 10).

*Theorem 4.* A fake cube is real if it has a triangulation that can be shelled.

*Figure 10*

**PROOF.** It can be shown inductively that each of  $s_n, (s_{n-1} + s_n), (s_{n-2} + s_{n-1} + s_n), \dots, (s_1 + s_2 + \dots + s_n)$  is a real cube.

Contrasted with the easiness with which any triangulation of a 2-cell can be shelled, there is difficulty in shelling a cube. It sometimes comes as a shock to the beginner that even real cubes have triangulations which cannot be shelled. Mary E. Rudin described [23] a rectilinear triangulation of a real tetrahedron that could not be shelled. Earlier unshellable triangulations of topological cubes were given in [9] and [27]. We describe two triangulations of a topological cube that cannot be shelled.

*Example 1.* Let  $C$  be the house with two rooms shown in Figure 11. The walls are made out of one layer of brick (without mortar) so that each brick touches air in two components, and if two bricks intersect, the intersection is a face, edge, or corner of each. The object is a house with two floors so that to enter the lower floor, one goes through a tunnel from the roof and to get to the upper floor one goes through a tunnel from below. There is a partition in each room from the wall to the tunnel leading to the other room. This causes the room to be simply connected.

We see that  $C$  is a topological cube if we regard it as a real cube with indentations from above and below.

Suppose  $C$  is triangulated as follows. Order the corners of the bricks as follows. First put in any vertex which is a component of the intersection of a brick with  $Bd\ C$ . There is one of these on each

inside corner of a room. Next, list the vertices not of this sort but which lie on an edge of a brick that is a component of the intersection of the brick with  $Bd C$ . Finally list the other vertices. The bricks are not triangulated as follows. A face of a brick is divided with a diagonal that contains the first vertex on the face. Once the faces of a brick are triangulated, the brick is triangulated by taking the cone from the first vertex of the brick to the triangles on the boundary of the brick that do not contain this vertex.

Note that each tetrahedra of the triangulation intersects the boundary of the cube in a disconnected set. Hence the triangulation cannot be shelled. It may be noted that each vertex of the triangulation lies on  $Bd C$ .

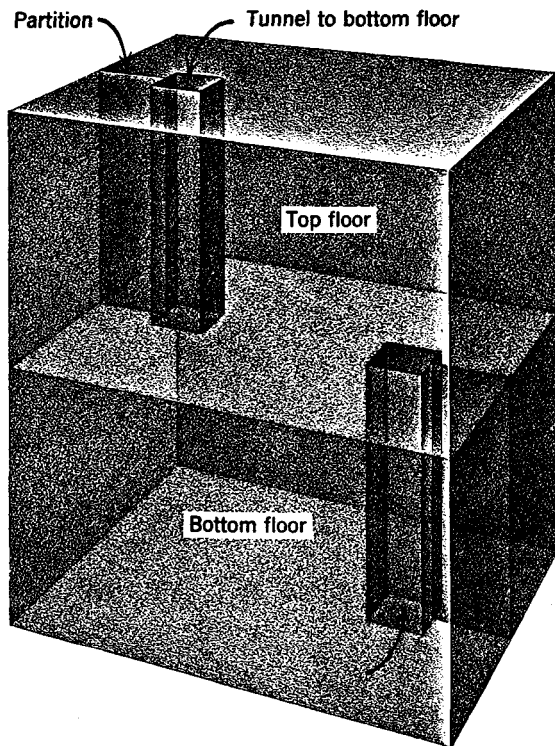
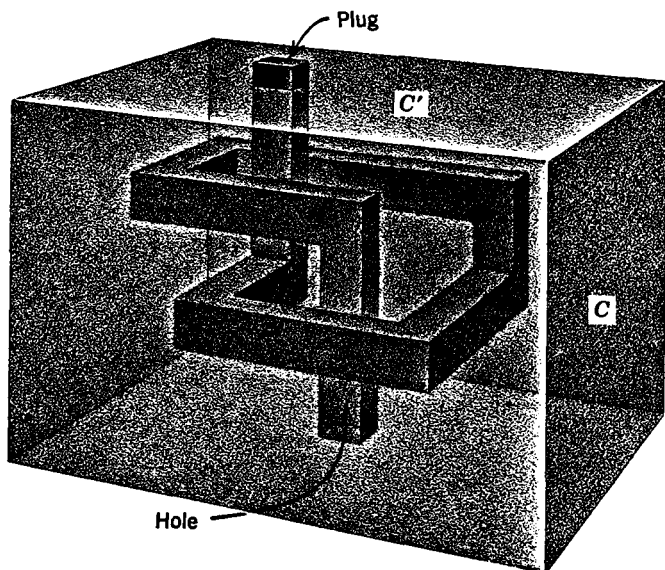


Figure 11

*Figure 12*

*Example 2.* Consider the cube with the plugged knotted hole shown in Figure 12. The object is topologically a cube since the hole is plugged at the upper end with a small cube  $C'$ . The resulting topological cube  $C$  is triangulated so that the edges of  $C'$  are  $l$ -simplexes of the triangulation. This may be accomplished by first subdividing  $C$  into small cubes of the same size of  $C'$ , subdividing each cube into two prisms, and then subdividing each prism into three tetrahedra—taking care that the resulting tetrahedra fit together right.

Consider a simple closed curve  $J$  made up of a spanning segment of  $C$  which is an edge of  $C'$  and an arc of  $Bd\ C$ . Then  $J$  is knotted. If we were to start shelling the triangulation of  $C$ , at each stage there would be a knotted simple closed curve which lies, except for one spanning simplex, on the boundary of the resulting cube. It follows that the triangulation of  $C$  cannot be shelled for at the last stage there could be no such simple closed curve.

Closely connected with the notion of shelling is that of sequential connectedness. A triangulation of a finite  $n$ -complex all of whose

simplexes are faces of an  $n$ -simplex is *sequentially connected* if its  $n$ -simplexes can be ordered  $s_1, s_2, \dots, s_n$  so that for each  $i < n$ ,  $s_i$  intersects  $(s_{i+1} + s_{i+2} + \dots + s_n)$  in a connected set. The notion of sequential connectedness is stronger than either that of connectedness or unicoherence since a projective plane is connected and unicoherent. It has no sequentially connected triangulation or else it would be simply connected.

The following theorem comes from [2].

**Theorem 5.** A fake cube is real if it has a sequentially connected triangulation.

A still more general sort of shelling is frequently used where simplexes of all dimensions are considered. A complex  $C$  is *collapsible* if it has a decreasing sequence of subcomplexes  $C = C_1, C_2, \dots, C_n$  such that  $C_n$  has only one vertex and for each  $i < n$ ,  $C_{i+1}$  is obtained from  $C_i$  by eliminating from  $C_i$  exactly two simplexes such that one is a major face of the other. If  $s, s'$  are the simplexes eliminated from  $C_i$ , with  $s'$  a face of  $s$ , it follows that  $s$  is not a face of any other simplex of  $C_i$ , and  $s'$  is not a face of any simplex except  $s$  and  $s'$  itself.

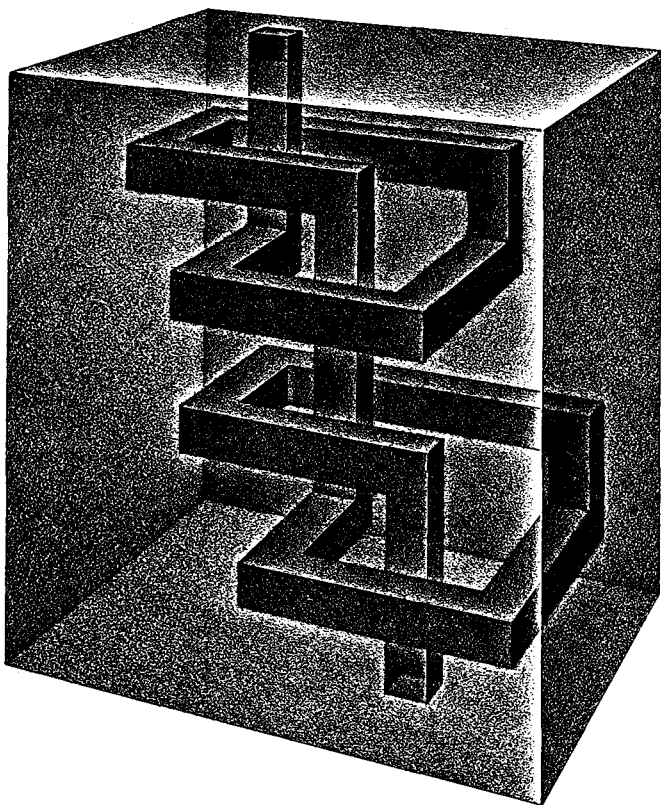
**Theorem 6.** A fake cube is real if it has a collapsible triangulation.

**PROOF.** The proof is found in works of H. C. Whitehead [32], but for completeness we indicate a proof in the following paragraph.

For each simplex  $s$  of the triangulation, let  $S(s)$  be the star of the baricenter of  $s$  under the second baricentric subdivision of the collapsible triangulation. Then these  $S(s)$ 's provide the fake cube with a cellular subdivision. By following the way that the original triangulation can be collapsed, we find that this cellular subdivision can be shelled. Then the proof follows that of Theorem 3.

It is not immediately obvious whether or not the triangulation given by Mary E. Rudin in [23] or that of a house with two rooms or that of a cube with a plugged knotted hole is either sequentially connected or collapsible. We give an example that is not.

**Example 3.** This example is the same as Example 2 except that the knot is more complicated and consists of a double trefoil knot instead of a single one (see Figure 13). The proof that the resulting triangulation is not collapsible is given elsewhere. By getting more complicated holes we can show that for each integer  $n$  there is a

*Figure 13*

triangulation of a cube whose  $n$ th barycentric subdivision is not contractible. We now consider another condition on the triangulation of a fake cube that will insure that it is real cube.

*Theorem 7.* A fake cube  $K$  is a cube if and only if it has a triangulation  $T$  such that each vertex of  $T$  lies on  $Bd K$  and each 3-simplex of  $T$  has at least two edges on  $Bd K$ .

**PROOF.** The necessity of the condition follows from the fact that a cube has a triangulation with only one 3-simplex. We now show that a fake cube is real if it has a triangulation satisfying the prescribed condition.

Following Theorem 2, we suppose with no loss of generality that each 2-simplex of  $T$  has precisely one or three edges on  $Bd K$ .

We show here that each 1-simplex of  $T$  lies on  $Bd K$ . Assume  $T$  has 1-simplexes  $s_1^1, s_2^1, \dots, s_m^1$  not on  $Bd K$ . The fundamental group of  $Bd K + \Sigma s_1^1$  is the free group on  $m$  generators

$$G = \{a_1, a_2, \dots, a_m\}$$

Adding a 2-simplex  $s^2$  of  $T$  with three edges on  $Bd K$  to  $Bd K + \Sigma s_i^1$  does not change the fundamental group of the resulting set since  $Bd s^2$  could already be shrunk to a point on  $Bd K$ . However, adding a 2-simplex  $s^2$  of  $T$  with one edge on  $Bd K$  to  $Bd K + \Sigma s_i^1$  results in a set whose fundamental group is a group with  $m$  generators and one relation  $w = 1$  where  $w$  is a two letter word (since  $s^2$  has precisely two edges among the  $s_i^1$ 's). Accordingly, the fundamental group of the 2-skeleton of  $T$  is

$$G' = \{a_1, a_2, \dots, a_n | w_1 = w_2 = \dots = w_n = 1\}$$

where each of the  $w_i$ 's is a two letter word. This group is non-trivial because the homomorphism  $f(a_i) = b$  sends  $G'$  onto the group  $H = \{b/b^2 = 1\}$ . This contradicts the fact that  $K$  is simply connected. This contradiction arises from assuming that  $T$  has a 1-simplex not on  $Bd K$ .

Since each 1-simplex of  $T$  lies on  $Bd K$ , each 2-simplex of  $T$  either lies on  $Bd K$  or separates  $K$ . Hence  $T$  can be shelled. Theorem 7 then follows from Theorem 4.

## 5. EULER CHARACTERISTICS?

For a triangulation of a complex, let  $V_i$  denote the number of  $i$ -simplexes in the triangulation. For triangulations of a compact connected orientable 2-manifold, the number

$$V_2 - V_1 + V_0$$

is important since it determines the topology of the 2-manifold. We might hope we could tell whether or not a fake cube is real by considering a triangulation and computing the  $V_i$ 's. We find the following, however.

**Theorem 8.** For each triangulation  $T$  of a compact 3-manifold,  $V_2 = 2V_3$ .

**PROOF.** We merely note that each 3-simplex of  $T$  has four faces and each face belongs to precisely two 3-simplexes.

Theorem 8 shows us that we do not need to concern ourselves with the number of 3-simplexes. This might seem temporarily encouraging, for whether or not a triangulated homotopy sphere is topologically  $S^3$  depends not on the number of 3-simplexes, but rather on whether or not the 2-skeleton of the triangulation can be embedded in  $E^3$ . We then find the following discouraging result, however.

**Theorem 9.** For each triangulation  $T$  of a compact 3-manifold,  $V_1 = V_3 + V_0$ .

**PROOF.** For each vertex  $v$  of  $T$ , let  $V_i(v)$  be the number of  $i$ -simplexes of  $T$  containing  $v$ . It follows from Euler's formula applied to the 2-sphere  $L(v, T) = Bd\ Star(v, T)$  that

$$V_3(v) - V_2(v) + V_1(v) = 2$$

Summing with respect to all the vertices of  $T$ , we find that

$$\Sigma[V_3(v) - V_2(v) + V_1(v)] = 2V_0$$

Each  $i$ -simplex contributes to  $i + 1$  of the  $V_i(v)$ 's so

$$\Sigma V_i(v) = (i + 1)V_i$$

Hence,

$$4V_3 - 3V_2 + 2V_1 = 2V_0$$

Theorem 9 follows by substituting the equation from Theorem 8 in the foregoing equation.

As a final discouraging factor, we have the following theorem.

**Theorem 10.** If  $M_1, M_2$  are two (possibly topologically different) compact 3-manifolds, then there are triangulations  $T_1, T_2$  of  $M_1, M_2$  respectively such that  $T_1, T_2$  have precisely the same number of simplexes in each dimension.

**PROOF.** I had a rather cumbersome proof of this result for orientable manifolds using an interesting theorem by Wallace [28] and Lickorish [10], but Leslie Glaser pointed out the following easy proof.

Let  $T'_1, T'_2$  be arbitrary triangulations of  $M_1, M_2$ . By subdividing one of them we get triangulations with the same number of

vertices. If the triangulations also have the same number of 3-simplexes, Theorems 8 and 9 imply that they have the same number of simplexes of each dimension.

For the one of the triangulations with the larger number of 3-simplexes, subdivide the triangulation by adding a vertex at the baricenter of a 3-simplex. This increases the number of vertices by one and the number of 3-simplexes by three. To the other, add a vertex at the baricenter of a 2-simplex. This increases the number of vertices by one and the number of 3-simplexes by four. Hence, in a finite number of such subdivisions, we bring  $T'_1$  and  $T'_2$  into balance with respect to the number of simplexes in each dimension (see Figure 14).

This result suggests that knowing the  $V_i$ 's does not give enough information. Although  $T_1$  and  $T_2$  have the same number of simplexes in each dimension, they may differ in how these simplexes fit together. Something might be accomplished by putting coefficients on the simplexes and weighting them according to how other simplexes hang on to them—but this would be more difficult than counting.

*Theorem 11.* For each triangulation of a compact 3-manifold,  $V_3 - V_2 + V_1 - V_0 = 0$ .

PROOF. The result is an immediate consequence of Theorems 8 and 9.

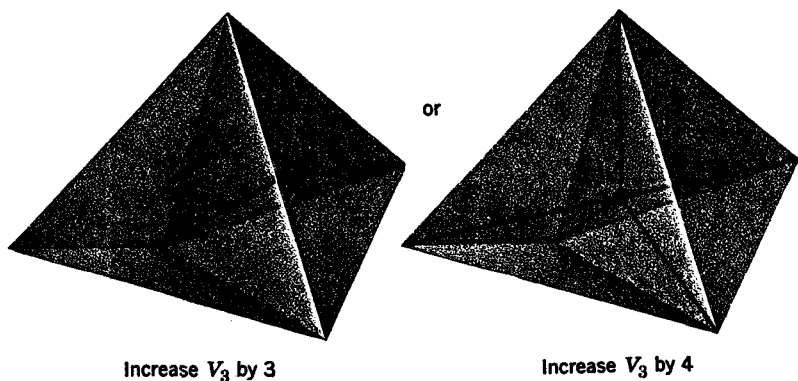


Figure 14

## 6. CURVES ENCIRCLED BY TORI

An affirmative answer to the following problems would provide a partial solution to the Poincaré conjecture.

*Question 1.* Is a compact connected 3-manifold  $M$  topologically  $S^3$  if each tame simple closed curve in  $M$  can be isotopically shrunk to small size in  $M$ ?

A simple closed curve  $J$  in  $M$  can be isotopically shrunk to small size in  $M$  if for each  $\epsilon > 0$  there is an isotopy  $h: J \times [0, 1] \rightarrow M$  such that  $h_0 = I$  (identity), and diameter  $h_1(J) < \epsilon$ .

Theorem 19 of Section 7 would provide an affirmative answer to Question 1 if we have an affirmative answer to the following question.

*Question 2.* Can each simple closed curve in a homotopy 3-manifold be encircled by a solid torus?

A *solid torus* is an object topologically equivalent to the cartesian product of a circle  $S^1$  with a disk  $D$ . If  $p_0 \in \text{Int } D$ , the image of  $S^1 \times p_0$  is called a *core* of the solid torus. A solid torus  $T$  *encircles* a simple closed curve  $J$  if  $J \subset \text{Int } T$  and  $J$  is homotopic in  $\text{Int } T$  to a core of  $T$ . In fact, for map  $f: J \rightarrow \text{Int } T$  we say that  $T$  encircles  $f(J)$  if there is a homeomorphism  $h: J \rightarrow \text{Int } T$  such that  $T$  encircles  $h(J)$  and  $f$  is homotopic to  $h$  in  $\text{Int } T$ .

Not every map of a simple closed curve into  $E^3$  can be encircled by a solid torus but it can if the map is a homeomorphism.

*Theorem 12.* Each simple closed curve in  $E^3$  is encircled by a solid torus.

**PROOF.** Let  $a, b$  be two points of a simple closed curve  $J$  and  $P$  be a plane separating  $a$  from  $b$ . Let  $axb$  be one arc from  $a$  to  $b$  in  $J$  and  $ayb$  be the other. There is a polygonal simple closed curve  $K$  in  $P$  separating  $axb \cdot P$  from  $ayb \cdot P$  in  $P$ . Let  $T$  be a tubular neighborhood of  $K$  so small that it misses  $J$ . If the point at infinity is added to  $E^3$ , the complement of  $T$  is a solid torus which encircles  $J$ .

To get a solid torus encircling  $J$  without adding the point at infinity, we could consider a large round ball  $B$  whose interior contains  $K + J$ . Let  $A$  be a polygonal arc in  $P - J$  from  $K$  to  $Bd B$ . Then the solid torus is  $B$  minus a tubular neighborhood of  $A + K$ .

Each polygonal simple closed curve  $J$  is encircled by a solid torus each point of which lies near  $J$ . We should not conclude that the torus promised by Theorem 12 lies close to the simple closed curve it encircles because the simple closed curve was not assumed to be polyhedral. Neither should we conclude that Theorem 12 holds in 3-manifolds other than  $E^3$  and  $S^3$ .

*Example 4.* There is a simple closed curve  $J$  in  $E^3$  such that each solid torus encircling it extends far away. Consider a wild 2-sphere  $S$  in  $E^3$  with only a 0-dimensional set of wild points such that the complement of  $S$  is not simply connected (see Figure 15). Let  $J$  be

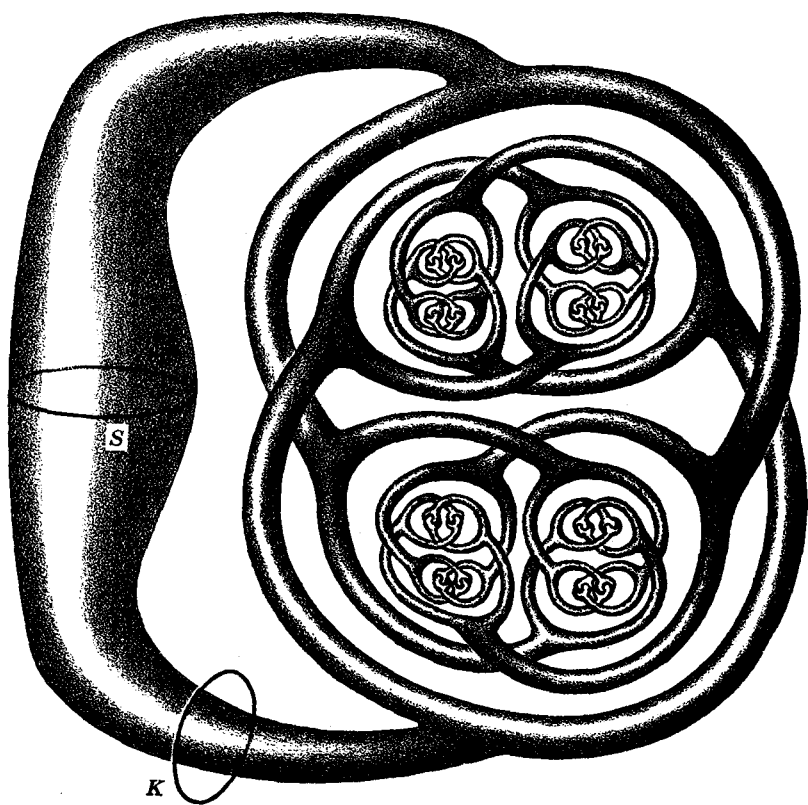


Figure 15

a simple closed curve on  $S$  which contains all the wild points of  $S$ , and  $K$  be a polygonal simple closed curve in  $E^3 - S$  which cannot be shrunk to a point in  $E^3 - S$ . Then  $K$  cannot be shrunk to a point in  $E^3 - J$  even though  $J$  bounds a disk in  $S \subset E^3 - K$ . We indicate in the following paragraph why each solid torus encircling  $J$  must intersect  $K$ .

If  $T$  were a solid torus in  $E^3 - K$  encircling  $J$ , it would have a tame core  $C$ . This tame core  $C$  could be shrunk to a point in  $E^3 - K$  so  $C$  would not homologically link  $K$ . Then  $K$  would not homologically link  $C$ . It follows from Theorem 15 that  $C$  is unknotted so  $K$  could be shrunk to a point in  $E^3 - C$ . If  $K$  could be shrunk to a point in  $E^3 - C$ , however, it could be shrunk in  $E^3 - \text{Int } T$  and hence in  $E^3 - J$ .

Suppose simple closed curve  $J$  is oriented as is the longitudinal direction around a solid torus  $T$ . In the next two theorems, for each map  $f$  of  $J$  into  $T$ , we let  $L(f)$  be the number of times  $f$  circles  $T$  longitudinally. To find  $L(f)$  we could let  $C$  be a core of  $T$  with length oriented as  $T$  and  $r$  be a retraction of  $T$  onto  $C$ . Then subdivide  $J$  into  $n$  small positively oriented arcs  $x_1x_2, x_2x_3, \dots, x_nx_1$  and let  $L(f)$  be the sum of the oriented distances along  $C$  from  $rf(x_i)$  to  $rf(x_{i+1})$  divided by the length of  $C$  where it is understood that we denote  $x_{n+1}$  by  $x_1$ .

**Theorem 13.** If  $f_1, f_2$  are maps of  $J$  into  $\text{Bd } T$  such that  $L(f_1) \geq 2$ ,  $L(f_2) \neq 0$ , then either  $L(f_2)$  is a multiple of  $L(f_1)$  or else there is a map  $f$  of  $J$  into  $f_2(J) + f_1(J)$  such that

$$0 < L(f) < L(f_1)$$

**PROOF.** We suppose  $L(f_2) > 0$ . Assume the theorem is false and that  $L(f_2)$  is minimal in the sense that if  $f_1, f_3$  are two other maps showing the theorem false; then  $L(f_2) \leq L(f_3)$ . Note that  $L(f_1) < L(f_2)$ .

First we consider the case where  $f_1(J), f_2(J)$  have a point  $p$  in common. Let  $f_3$  be a map of  $J$  onto  $f_1(J) + f_2(J)$  so that as a point moves half way around  $J$ , its image under  $f_3$  starts at  $p$  and traces out  $f_2(J)$ , and as the point goes the other half way around  $J$ , the image under  $f_3$  traces out  $f_1(J)$  in the negative direction so that

$$L(f_3) = L(f_2) - L(f_1)$$

But  $L(f_3)$  is not a multiple of  $L(f_1)$  unless  $L(f_2)$  is and  $L(f_3) < L(f_2)$ . This contradicts the minimality we had assumed for  $L(f_2)$ .

Finally we consider the case where  $f_1(J) \cdot f_2(J) = 0$ . Let  $f'_i$  ( $i = 1, 2$ ) be a homeomorphism of  $J$  into  $f_i(J)$  such that  $L(f'_i) > 0$ . Since  $f'_1(J) \cdot f'_2(J) = 0$ ,

$$L(f'_1) = L(f'_2)$$

Since  $f_1(J) \cdot f'_2(J) = 0$ ,  $L(f_1)$  is a multiple of  $L(f'_2)$  and hence a multiple of  $L(f'_1)$ . Then  $L(f'_1) \leq L(f_1)$ , and under the assumption that there is no  $f$  with  $0 < L(f) < L(f_1)$  we have

$$L(f_1) = L(f'_1)$$

Since  $f_2(J) \cdot f'_1(J) = 0$ ,  $L(f_2)$  is a multiple of  $L(f'_1)$  and hence a multiple of  $L(f_1)$ .

A *punctured disk* is a set homeomorphic to  $D_0 - (\text{Int } D_1 + \text{Int } D_2 + \cdots + \text{Int } D_n)$  where  $D_0$  is a disk and  $D_1, D_2, \dots, D_n$  is a collection of mutually exclusive disks each lying in  $\text{Int } D_0$ .

**Theorem 14.** Suppose  $f$  is a map of a punctured disk  $D_0 - (\text{Int } D_1 + \text{Int } D_2 + \cdots + \text{Int } D_n)$  into a solid torus  $T$  such that  $L(f/Bd D_0) = 1$  and  $f(Bd D_1 + Bd D_2 + \cdots + Bd D_n) \subset Bd T$ . Then each open subset of  $Bd T$  containing  $f(Bd D_1 + Bd D_2 + \cdots + Bd D_n)$  also contains a simple closed curve isotopic in  $T$  to a core of  $T$ .

**PROOF.** Suppose each  $Bd D_i$  ( $i = 1, 2, \dots, n$ ) is ordered opposite that of  $Bd D_0$ . By joining the  $Bd D_i$ 's with cross cuts in the punctured disk, we have,

$$L(f|Bd D_0) = L(f|Bd D_1) + L(f|Bd D_2) + \cdots + L(f|Bd D_n) = 1$$

Let  $n_0$  be a positive integer minimal with respect to there being a map  $g$  of  $S^1$  into  $f(Bd D_1 + Bd D_2 + \cdots + Bd D_n)$  such that  $L(g) = n_0$ . Either  $n_0 = 1$  or else by the foregoing relation there is an  $L(f/Bd D_i)$  which is not a multiple of  $n_0$ . Theorem 13 shows that this second alternative leads to the contradiction that  $n_0$  is not minimal. Hence  $n_0 = 1$ .

Let  $U$  be an open subset of  $Bd T$  containing  $g(S^1)$  and such that each boundary component of  $U$  is a simple closed curve. We show that  $U$  contains a simple closed curve that is isotopic in  $T$  to a core of  $T$ .

If each boundary component of  $U$  bounds a disk in  $T$ , Theorem 14 follows since we can get a simple closed curve  $C$  in  $Bd\ T$  which circles  $T$  longitudinally once and then push this  $C$  off the disks of  $Bd\ T - U$ . If  $J$  is a boundary component of  $U$  which does not bound a disk,  $J$  must circle  $T$  longitudinally, for if it circles it meridianally but not longitudinally,  $g(S^1)$  could not circle longitudinally, since  $g(S^1)$  misses  $J$ . In fact,  $L(g)$  is a multiple of the number of times  $J$  circles longitudinally. Since  $L(g) = 1$ ,  $J$  circles  $T$  once longitudinally. A simple closed curve  $J'$  in  $U$  runs "parallel" to  $J$  and circles  $T$  once longitudinally also. Hence  $J'$  is isotopic in  $T$  to a core of  $T$ .

It may be noted that we could not conclude that there is a simple closed curve in  $f(Bd\ D_1 + Bd\ D_2 + \cdots + Bd\ D_n)$  isotopic to a core of  $T$  because it might be that  $n = 2$ ,  $f(Bd\ D_1)$  is a simple closed curve circling  $T$  twice and  $f(Bd\ D_2)$  is a simple closed curve circling  $T$  three times with  $f(Bd\ D_1) \cdot f(Bd\ D_2)$  being one point.

**Theorem 15.** If  $D$  is a disk in  $E^3$  and  $T$  is a solid torus in  $E^3$  encircling  $Bd\ D$ , each core of  $T$  is unknotted.

**PROOF.** With no loss of generality, we suppose  $Bd\ T$  is polyhedral. To see that there is no loss of generality in supposing this we argue as follows. Suppose  $T$  is the image of a polyhedral torus  $T'$  under a homeomorphism  $h$ . It follows from [6] or [15] that we can adjust  $h$  slightly to get a homeomorphism  $h'$  of  $T'$  onto  $T$  such that  $h'$  is locally piecewise linear mod  $Bd\ T'$ . Reduce  $T'$  slightly to a polyhedral torus  $T''$  so that  $T''$  has the same core as  $T'$ ,  $T' \subset \text{Int}\ T''$ , and  $Bd\ D \subset h'(\text{Int}\ T'')$ . Then  $h'(T'')$  is a polyhedral torus with the same core as  $T$ .

We suppose that  $D$  is locally polyhedral mod  $Bd\ D$ , and from [4] we can support this. We suppose furthermore that  $D \cdot Bd\ T$  is the sum of a finite number of mutually exclusive simple closed curves. Such a desirable condition results from a slight adjustment of either the vertices of  $Bd\ T$  or of certain vertices on  $Bd\ D$ .

If a component of  $D \cdot Bd\ T$  bounds a disk on  $Bd\ T$ , we take the inside of such a disk, replace a disk in  $D$  with it, and shove it to one side of  $T$  so as to eliminate a component of  $D \cdot Bd\ T$ . For convenience we call the new disk  $D$  also. This elimination of components of  $D \cdot Bd\ T$  is continued until no component of  $D \cdot Bd\ T$  bounds a disk on  $Bd\ T$ .

It follows from Theorem 14 that each simple closed curve in

$D \cdot Bd T$  circles  $T$  longitudinally exactly once. Let  $D'$  be a subdisk of  $D$  such that  $D' \cdot Bd T = Bd D'$ . A core  $C$  of  $T$  is unknotted since it can be homotopically pulled in its own complement over to  $Bd D'$  and can then be shrunk to a point in  $D'$ . A tame simple closed curve is unknotted if it can be shrunk to a point in its own complement.

## 7. NICE SHRINKING

If  $J$  is a simple closed curve in  $E^3$ , then  $J$  can be shrunk to a point so that at each stage of the shrinking, except the last,  $J$  has been shrunk to a simple closed curve. A natural way to get this shrinking is to pick a point  $p$  of  $E^3$  and decide that at time  $t$  the point  $x$  of  $J$  has been moved to the point of the segment  $xp$  that divides it in the ratio  $t$  to  $(1 - t)$ . Note that if  $J$  is a polygonal simple closed curve, then at each stage of the homotopy except the last,  $J$  is a polygonal simple closed curve. It is not known whether or not this property holds in an arbitrary simply connected compact 3-manifold, but Theorem 17 shows that unless the Poincaré conjecture is true, it does not.

In this section we lean heavily on the following result proved in [5]. McMillan [13] has an even stronger version of this result.

*Theorem 16.* A compact, connected 3-manifold  $M$  is topologically  $S^3$  if each tame simple closed curve in  $M$  lies in a topological cube in  $M$ .

PROOF. Although we do not repeat the proof given in detail in [5], we list the steps used in proving the theorem.

1. Let  $T$  be a triangulation of  $M$ . We suppose each polygonal simple closed curve in this triangulation lies in a topological cube in  $M$ .

2. Show that step 1 implies that each polygonal simple closed curve lies on the interior of a polyhedral cube.

3. Show that step 2 implies that the 1-skeleton of  $T$  lies on the interior of a polyhedral cube.

4. Show that step 3 implies that the 2-skeleton of  $T$  lies in a polyhedral punctured cube.

5. Fill in the holes of the punctured cube of step 4 to show that  $M$  is topologically  $S^3$ .

The following result is new and its proof is included.

**Theorem 17.** A compact connected 3-manifold  $M$  is topologically  $S^3$  if for each tame simple closed curve  $J$  in  $M$  there is a homotopy  $f_t: J \rightarrow M$  ( $0 \leq t \leq 1$ ) such that  $f_0$  is the identity, each  $f_t(J)$  is encircled by a solid torus in  $M$ , and  $f_1(J)$  lies on the interior of a topological cube in  $M$ .

**PROOF.** We prove Theorem 17 by showing that each simple closed curve in  $M$  lies on the interior of a topological cube in  $M$ . Then Theorem 17 follows from Theorem 16.

Let  $X$  be the set of all points  $t$  of  $[0, 1]$  such that  $f_t(J)$  lies on the interior of a cube in  $M$ . The proof is finished if we show that  $0 \in X$ . It follows easily that  $X$  is an open subset of  $[0, 1]$  that contains 1. It follows that  $0 \in X$  when we show that  $X$  is closed. Toward this end we let  $t_0$  be a point of the closure of  $X$  and show that  $t_0 \in X$ .

Let  $T_0$  be a solid torus in  $M$  encircling  $f_{t_0}(J)$ . We select such a  $T_0$  whose boundary is polyhedral. That we can do this follows from an argument used in the proof of Theorem 15.

If there is a cube  $C$  in  $M$  whose interior contains a simple closed curve on  $Bd T_0$  that circles  $T_0$  once longitudinally, there is a cube  $C'$  whose interior contains a core  $K$  of  $T_0$ . Expanding a small tubular neighborhood of  $K$  to contain  $f_{t_0}(J)$ , we see how to adjust  $C'$  so that its interior contains  $f_{t_0}(J)$ . Hence we finish the proof of Theorem 17 by showing that there is a cube  $C$  in  $M$  such that  $\text{Int } C$  contains a simple closed curve on  $Bd T_0$  that circles  $T_0$  once longitudinally.

Let  $t_1$  be a point of  $X$  so near  $t_0$  that  $T_0$  encircles  $f_{t_1}(J)$  and  $C$  be a topological cube in  $M$  whose interior contains  $f_{t_1}(J)$ . Let  $D$  be a polyhedral disk and  $h: J \rightarrow Bd D$  be a homeomorphism. Let  $g: D \rightarrow \text{Int } C$  be a map such that  $g = f_{t_1}h^{-1}$  on  $Bd D$  and  $g$  is piecewise linear in a neighborhood of  $g^{-1}[Bd T_0 \cdot g(D)]$ . We suppose that  $g$  is in a general position with respect to  $Bd T_0$  in the sense that  $g^{-1}[Bd T_0 \cdot g(D)]$  is the sum of a finite collection of mutually exclusive simple closed curves. Let  $E$  be the component of  $D - g^{-1}[Bd T_0 \cdot g(D)]$  containing  $Bd D$ . Then  $\bar{E}$  is a punctured disk, and it follows from Theorem 14 that there is a simple closed curve in  $Bd T_0 \cdot \text{Int } C$  that is isotopic in  $T_0$  to a core of  $T_0$ .

We have the following as a corollary to Theorem 17.

**Theorem 18.** A compact connected 3-manifold  $M$  is topologically  $S^3$  if each tame simple closed curve  $J$  in  $M$  can be isotopically shrunk

to small size in  $M$  so that the image of  $J$  at each stage of the isotopy is a tame simple closed curve.

Although it is not known that each tame simple closed curve  $J$  in a homotopy 3-sphere can be shrunk to a point so that at each intermediate stage the image is a simple closed curve, it is known that  $J$  can be shrunk so that at each intermediate stage the image of  $J$  is either a simple closed curve or a figure eight.

**Theorem 19.** If  $J$  is a polygonal simple closed curve which can be shrunk to a point in a triangulated 3-manifold  $M$ , there is a homotopy  $f_t: J \rightarrow M$  such that  $f_0$  is the identity,  $f_1$  is a constant map, and each  $f_t$  ( $0 \leq t < 1$ ) is a piecewise linear map which is a homeomorphism except at a finite number of  $t$ 's and has precisely two singularities at each of these.

**PROOF.** Let  $D$  be a 2-simplex and  $g: D \rightarrow M$  be a piecewise linear map such that  $g: \text{Bd } D \rightarrow J$  is a homeomorphism and  $g$  is a general position map in the sense that there is a triangulation  $T$  of  $D$  such that  $g$  is a piecewise linear homeomorphism on each simplex of  $T$ ; and if  $s', s''$  are two simplexes of  $T$ ,

$$\text{dimension } [g(s') \cdot g(s'') - g(s' \cdot s'')] \leq \text{dimension } s' + \text{dimension } s'' - 3$$

Since  $T$  can be shelled (Theorem 3), there is an ordering  $s_1, s_2, \dots, s_n$  of the 2-simplexes of  $T$  such that for each  $i \leq n$ ,  $s_i + s_{i+1} + \dots + s_n$  is a disk  $D_i$ . Note that  $g$  is a homeomorphism on each  $\text{Bd } D_i$  and  $g^{-1}$  is locally a homeomorphism modulo a finite set of points  $X_i$  on  $g(s_i + \text{Bd } D_{i+1})$  where each point of  $X_i$  has precisely two preimages under  $g$ .

The homotopy  $f_t: J \rightarrow g(D)$  is described as follows.

$$f_0 = I$$

$$f_{i/n}(J) = g(\text{Bd } D_{i+1})$$

$$f_t = f_{i/n} \text{ on } f_{i/n}^{-1}g(\text{Bd } D_{i+1} \cdot \text{Bd } D_{i+2})$$

$$f_t = gh_tg^{-1}f_{i/n} \text{ on } f_{i/n}^{-1}g(\text{Bd } D_{i+1} - \text{Bd } D_{i+2})$$

for  $i/n \leq t \leq (i+1)/n$  where  $h_t$  is a piecewise linear isotopy that shoves  $\text{Bd } D_{i+1} - \text{Bd } D_{i+2}$  straight to  $\text{Bd } D_{i+2} - \text{Bd } D_{i+1}$  through  $s_{i+1}$  in such a way that for no  $t$  does  $gh_t(\text{Bd } D_{i+2} - \text{Bd } D_{i+1})$  con-

tain more than one point of  $X_{i+1}$  and for each point  $x$  of  $s_{i+1} \cdot X_{i+1}$  there is exactly one  $h_i$  with  $x \in gh_i(Bd D_{i+1} - Bd D_{i+2})$ . For the values of  $t$  for which  $gh_t(Bd D_{i+1} - Bd D_{i+2})$  misses  $X_{i+1}$ ,  $f_t$  is a homeomorphism and for the finite number of others,  $f_t(J)$  is a figure eight.

## 8. WHY?

Why are mathematicians interested in finding a solution to the Poincaré conjecture? For one thing, a solution would show the power of the algebraic technique. For compact 3-manifolds, it would show at least in one case that the algebraic properties determine the topology. It would also provide us with a handy characterization of  $S^3$ .

Aside from this, there is a historical interest in the Poincaré conjecture. Poincaré announced in [21] that two compact combinatorial  $n$ -manifolds were topologically equivalent if they had the same homological characters, but he disproved this in [22] by constructing a 3-manifold which is not simply connected, even though all its 0, 1, and 2 cycles bounded. The homology spheres given in Section 2 were such examples. Alexander showed [1] that there were combinatorially different toroidal 3-manifolds which had the same homotopy properties. It follows as a result of Moise's proof of the Hauptvermutung [16] that these homotopically equivalent toroidal manifolds were topologically different. We might still hope, however, that if the first homotopy group is trivial, this would imply that the manifolds were topologically equivalent also. J. H. C. Whitehead gave a faulty proof of the Poincaré conjecture in [30] but pointed out his error shortly thereafter in [31].

There were two competing solutions for the Poincaré conjecture being rumored at the Institute for Advanced Study in Princeton in 1957. Although both proofs fell flat, the rumor was so strong for awhile that the author considered withdrawing his partial solution [5] which had already been accepted by the *Annals*.

A "proof" of over 100 pages appeared [12] in 1958. The last half of the paper gives the true result that a compact connected triangulated 3-manifold is topologically  $S^3$  if its 2-skeleton can be embedded in  $E^3$ . The first half asserts that for any triangulation of a homotopy 3-sphere, the 2-skeleton can be embedded in  $E^3$ . The plan of the proof is to embed the 2-simplexes one at a time, taking

care that the images fan out around edges and vertices in the way the original simplexes do. Most of the first half is clear where it is shown how to embed the first few 2-simplexes. It is then claimed that if an obstruction arises, it would have been possible to start over and by a different approach embed more of the 2-simplexes. The argument here is not clear.

It was shown by Stallings, Zeeman, Smale, and Wallace that in high dimensions certain other versions of the Poincaré conjecture hold [25, 33, 34, 24, 29]. Theorem 20 was shown.

*Theorem 20.* If  $M^n$  ( $n \geq 5$ ) is a compact combinatorial  $n$ -manifold that has the same connectedness properties as  $S^n$ , then  $M^n$  is topologically equivalent to  $S^n$ .

PROOF. Although we do not include details, the proof may be given with the following steps.

1. Using an engulfing lemma [26], show that any  $(n - 3)$ -dimensional polyhedron in  $M^n$  lies on the interior of a polyhedral  $n$ -cell in  $M^n$ .
2. Put the  $(n - 3)$ -skeleton  $K^{n-3}$  of the triangulation  $T$  of  $M^n$  on the interior of a topological  $n$ -cell  $C_1$ .
3. Get the dual of  $K^{n-3}$  by taking the first barycentric subdivision of  $T$  and letting the dual  $H^2$  be the sum of all simplexes of this subdivision that miss  $K^{n-3}$ . It is to be noted that  $H^2$  is of dimension 2.
4. If  $2 \leq n - 3$  (and this occurs only when  $n \geq 5$ ), there is a topological  $n$ -cell  $C_2$  whose interior covers  $H^2$ .
5. Swell up  $C_1$  and  $C_2$  until the sum of their interiors cover  $M^n$ .
6. It follows from the Schoenflies theorem proved by Brown [7] that any compact  $n$ -manifold is topologically  $S^n$  if it is the sum of two open  $n$ -cells.

One of the reasons we have been able to prove so many interesting things about 3-manifolds is that we know so much about  $S^2$ . Not knowing more about  $S^3$  is a bottleneck in learning more about higher dimensional manifolds. The following theorem gives one reason for wanting to get an answer to the Poincaré conjecture.

*Theorem 21.* Any triangulated 4-manifold is a combinatorial 4-manifold if the Poincaré conjecture holds.

This result follows from the following theorem.

**Theorem 22.** If  $T$  is a triangulation of a 4-manifold  $M^4$  and  $v$  is a vertex of  $T$ , then  $L(v, T) = \text{Bd Star}(v, t)$  is a simply connected 3-manifold.

**PROOF.** Note that  $L(v, T)$  consists of the sum of all 3-simplexes  $s^3$  of  $T$  such that the join of  $v$  and  $s^3$  is a 4-simplex of  $T$ . Let  $T'$  be the set of simplexes of  $T$  in  $L(v, T)$ . We note that  $L(v, T)$  is simply connected. If it were not, some simple closed curve near  $v$  in  $M^4 - v$  could not be shrunk to a point on a small subset of  $M^4 - v$ .

To show that  $L(v, T)$  is a manifold, it suffices to show that the link  $L(v', T')$  of each vertex  $v'$  of  $T'$  is a 2-sphere. This link  $L(v', T')$  is the sum of all 2-simplexes  $s^2$  of  $T'$  such that the join of  $s^2$  and  $v'$  is a 3-simplex of  $T'$ . We note that  $L(v', T')$  is a 2-complex. Also,  $L(v', T')$  is connected or else  $v'$  locally separates  $L(v', T')$  and  $vv'$  locally separates  $M^4$ . Each 1-simplex  $s^1$  of  $T'$  in  $L(v', T')$  is a face of precisely two 2-simplexes of  $T'$  in  $L(v', T')$  or else the invariance of domain is violated in  $M^4$  near the join of  $s^1$  and  $v$ . No point  $s$  of  $L(v', T')$  locally separates  $L(v', T')$  or else  $v$ 's locally separates  $L(v, T)$  and  $vv$ 's locally separates  $M^4$ . Hence  $L(v', T')$  is a connected 2-manifold.

To show that the connected 2-manifold  $L(v', T')$  is a 2-sphere, it suffices to show that each simple closed curve in  $L(v', T')$  homologically bounds in  $L(v', T')$ . If such a simple closed curve  $J_1$  did not bound, there would be a small simple closed curve  $J_2$  in  $L(v, T) - v'$  near  $v'$  such that  $J_2$  does not bound on a small subset of  $L(v, T) - v'$ . Then there would be a small simple closed curve  $J_3$  on the join of  $v$  and  $J_2$  near the center of  $vv'$  such that  $J_3$  does not bound on a small set in  $M^4 - vv'$ . This violates the fact that the first homology of a 4-cell remains trivial even on the deletion of a closed set homeomorphic to a subset of an arc.

## REFERENCES

1. J. W. Alexander, "Note on two three-dimensional manifolds with the same group," *Trans. Amer. Math. Soc.* **20** (1919) 339-342.
2. R. H. Bing, "A characterization of 3-space by partitioning," *Trans. Amer. Math. Soc.* **70** (1951) 15-27.
3. ———, "Locally tame sets are tame," *Ann. of Math.* **59** (1954) 145-158.
4. ———, "Approximating surfaces with polyhedral ones," *Ann. of Math.* **65** (1957) 454-483.
5. ———, "Necessary and sufficient conditions that a 3-manifold be  $S^3$ ," *Ann. of Math.* **68** (1958) 17-37.

6. ———, "An alternative proof that 3-manifolds can be triangulated," *Ann. of Math.* **69** (1959) 37–65.
7. M. Brown, "A proof of the generalized Schoenflies theorem," *Bull. Amer. Math. Soc.* **66** (1960) 74–76.
8. R. H. Fox, *Construction of Simply Connected 3-Manifolds, Topology of 3-Manifolds and Related Topics*, Englewood Cliffs, N.J. (1962) Prentice-Hall, 213–216.
9. F. Frankl, "Zur Topologie des dreidimensionalen Raumes," *Monatsh. Math. Phys.* **38** (1931) 357–364.
10. J. Hempel, *Construction of Orientable 3-Manifolds, Topology of 3-Manifolds and Related Topics*, Englewood Cliffs, N.J. (1962) Prentice-Hall, 207–212.
11. ———, "A simply connected 3-manifold is  $S^3$  if it is the sum of a solid torus and the complement of a torus knot," *Bull. Amer. Math. Soc.* **15** (1964) 154–158.
12. K. Koseki, "Poincarésche Vermutung in Topologie," *Math. J. Okayama Univ.* **8** (1958) 1–106.
13. D. R. McMillan, Jr., "On homologically trivial 3-manifolds," *Trans. Amer. Math. Soc.* **98** (1961) 350–367.
14. J. Milnor, "A unique decomposition of 3-manifolds," *Amer. J. Math.* **84** (1962) 1–7.
15. E. E. Moise, "Affine structures in 3-manifolds, IV. Piecewise linear approximations of homeomorphisms," *Ann. of Math.* **55** (1952) 215–222.
16. ———, "Affine structures in 3-manifolds, V. The triangulation theorem and Hauptvermutung," *Ann. of Math.* **56** (1952) 96–114.
17. C. D. Papakyriakopoulos, "On solid tori," *Proc. London Math. Soc.* (3) **7** (1957) 281–299.
18. ———, "On Dehn's lemma and the asphericity of knots," *Ann. of Math.* **66** (1957) 1–26.
19. ———, "Some problems on 3-dimensional manifolds," *Bull. Amer. Math. Soc.* **64** (1958) 311–335.
20. ———, "Reduction of the Poincaré conjecture to other conjectures," *Bull. Amer. Math. Soc.* **68** (1962) 360–366.
21. H. Poincaré, "Second complément à l'analysis situs," *Proc. London Math. Soc.* **32** (1900) 277–308.
22. ———, "Cinquième complément à l'analysis situs," *Rend. Circ. Mat. Palermo* **18** (1904) 45–110.
23. Mary E. Rudin, "An unshellable triangulation of a tetrahedron," *Bull. Amer. Math. Soc.* **64** (1958) 90–91.
24. S. Smale, "Generalized Poincaré's conjecture in dimensions greater than four," *Ann. of Math.* (2) **74** (1961) 391–406.
25. J. R. Stallings, "Polyhedral homotopy spheres," *Bull. Amer. Math. Soc.* **66** (1960) 485–488.
26. ———, "The piecewise-linear structure of euclidean space," *Proc. Cambridge Philos. Soc.* **58** (1962) 481–488.
27. E. R. Van Kampen, Remark on the address of S. S. Cairns, Michigan Lectures in Topology (1941).
28. A. H. Wallace, "Modifications and cobounding manifolds," *Canad. J. Math.* **12** (1960) 503–528.

29. A. H. Wallace, "Modifications and cobounding manifolds," II, *J. Math. Mech.* **10** (1961) 773-809.
30. J. H. C. Whitehead, "Certain theorems about three-dimensional manifolds," I, *Quart. J. Math. Oxford* **5** (1934) 308-320.
31. ———, "Three-dimensional manifolds (corrigendum)," *Quart. J. Math. Oxford* **6** (1935) 80.
32. ———, "Simplicial spaces, nuclei, and m-groups," *Proc. London Math. Soc.* **45** (1939) 243-327.
33. E. C. Zeeman, "The generalized Poincaré conjecture," *Bull. Amer. Math. Soc.* **67** (1961) 270.
34. ———, *The Poincaré conjecture for  $n \geq 5$ , Topology of 3-manifolds and related topics*, Englewood Cliffs, N.J. (1962) Prentice-Hall, 198-204.

# 4

## Partial Differential Equations: problems and uniformization in Cauchy's problem

Lars Gårding

### *Lecture I Some Problems in the Theory of Partial Differential Equations*

#### 1. INTRODUCTION

The theory of general partial differential operators has made big strides in the last ten or fifteen years. It has been treated in more or less comprehensive expositions [1, 2] and has now reached the monograph stage [3]. This lecture describes some recent developments for elliptic and hyperbolic equations and states some problems that they pose. There is also a concluding section on Frobenius' theorem and invariant distributions.

*Elliptic Operators.* Let  $X$  be an infinitely differentiable manifold of dimension  $n$ . Let

$$a(x, D) = \sum_{|\alpha| \leq m} a_\alpha(x) D^\alpha$$

where

$$D_k = \frac{1}{i} \frac{\partial}{\partial x_k}, \quad D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}, \quad |\alpha| = \alpha_1 + \cdots + \alpha_n$$

be a differential operator of order  $m$  with smooth complex coeffi-

cients defined in an open bounded part  $V$  of  $X$  with a smooth boundaries. Let

$$g(x, D) = \sum_{|\alpha|=m} a_\alpha(x) D^\alpha$$

be the principal part of  $a$ . Let  $B_1, \dots, B_l$  be similar operators of orders  $m_1, \dots, m_l$  defined on  $S$ . The classical linear boundary problems for Laplace's operator are of the form

$$(1.1) \quad au = v \text{ in } V, \quad B_j u = w_j \text{ on } S$$

The general theory of such problems (see, e.g., [4] and its bibliography) is based on a principle of localization due to Lopatinski [5], which can be formulated loosely as follows: the problem (1.1) is reasonable provided it is sufficiently reasonable at every point  $y \in S$  when we freeze the operators  $a$  and  $B_j$  at  $y$  [replace  $a(x, D)$  by  $a(y, D)$  and similarly for  $B_j$ ] and replace  $S$  by its tangent plane at  $y$  and  $V$  by a half-space that the plane bounds. Having done this we can also get rid of all the variables  $x$  except one by performing a Fourier transform in the tangent plane. Thus (1.1) is reduced to a one-dimensional problem:

$$(1.2) \quad a(y, p + ND_t)u(t) = v(t), \quad B_j(y, p + ND_t)u(0) = w_j$$

where  $N = N(y)$  is the inferior normal of  $S$  at  $y$ ,  $p = (p_1, \dots, p_n)$  is real,  $t \geq 0$  is a real variable, and  $D_t = -id/dt$ . Further, some experience with Fourier transforms will tell us that it is sufficient to consider (1.2) when  $p$  is large modulo  $N$ . That (1.2) is reasonable means that it has a unique solution which is small for large  $t$  provided  $v \neq 0$ . In general,  $g(x, N) \neq 0$ , so that we can factorize

$$a(x, p + ND_t) = g(x, N) \prod_{k=1}^m (D_t - \tau_k)$$

with

$$\tau_k = \tau_k(x, p, N)$$

When  $v = 0$ , any solution of (1.2) is a linear combination of exponential solutions

$$e^{i\tau_k t}, \quad \tau_k = \tau_k(y)$$

which fall off for large  $t$  if and only if  $\text{Im} \tau_k > 0$ . If (1.2) is to determine  $u$ , the number of such  $\tau_k$  ought to be the same as the number of linearly independent  $B$ 's. Trouble can be expected at points  $y$  where  $g(y, N) = 0$  since some  $\tau_k$  becomes infinite there.

We avoid this by requiring that

$$(1.3) \quad g(x, p) \neq 0 \text{ when } p \neq 0, \text{ real, and } x \in V + S$$

This is the well-known condition of ellipticity; it implies that all solutions of  $au = 0$  in  $V$  are smooth. Returning to the  $\tau_k$ , we observe that, because of (1.3),

$$\frac{a(x, p)}{g(x, p)} \rightarrow 1$$

as  $p \rightarrow \infty$ . This means that for large real  $p$ , the zeros of  $a(x, p + \tau N) = 0$  and  $g(x, p + \tau N) = 0$  are close of order  $o(|p|)$  and that the order of magnitude of the latter ones, which are homogeneous of order 1 in  $p$ , is  $|p|$ . Hence we can replace our condition on the zeros  $\tau$  by the following.

$$(1.4) \quad g(x, p + \tau N) = 0 \text{ has precisely } l \text{ zeros with positive imaginary parts for } x \in S \text{ and } p \text{ real and } \neq 0 \bmod N$$

and a loosely formulated condition:

$$(1.5) \quad \text{the operators } B \text{ are linearly independent}$$

When  $n > 2$ , we can pass from  $p$  to  $-p$  continuously, keeping  $p$  large modulo  $N$ . Hence (1.4) means that, necessarily,  $2l = m$  in this case. The simplest boundary problem is Dirichlet's problem where  $B_1, \dots, B_l$  are the first  $l$  normal derivatives. The general theory associated with (1.3), (1.4), and a quantitative form of (1.5) gives as a result that the mapping

$$(1.6) \quad u \rightarrow Tu = au, \quad B_1 u, \dots, B_l u$$

when defined on suitable function spaces has a null-space of finite dimension  $\nu$  and a range of finite co-dimension  $\rho$ . Increasing smoothness of  $Tu$  implies increasing smoothness of  $u$ . This theory covers, for example, the problem of the oblique derivative (a Laplace's operator,  $B = B_1$ , an arbitrary real first-order derivative) and has given a very interesting perspective on the classical boundary problems. But it also raises a number of problems.

In the first place, the index  $\nu - \rho$  becomes invariant under continuous deformations of the operator  $T$ . For the problem of the oblique derivative in two dimensions, it was discovered long ago by Fritz Noether [6] that the index equals twice the ordinary winding

number of the vector field  $B$  relative to the boundary  $S$ . Are there similar interpolations in more general cases? More generally, what is the connection between the index and the deformation classes? This question also has wider implications. The question of the index for an elliptic system on a manifold has recently begun to attract attention [7, 8]. Second, it is essential for the principle of localization that  $S$  be smooth. Can it be extended to cases when  $S$  has corners of a simple type? In that situation we can expect a complication: increased smoothness of  $Tu$  will not make  $u$  smooth. Consistency conditions at the corners are also required. Which are they and for which degrees of smoothness do they enter? How is the index influenced by the consistency conditions? For discontinuous boundary operators  $B$ , these problems have been studied in a very interesting paper by Peetre [9].

## 2. SOBOLEV SPACES AND BOUNDARY PROBLEMS

Let  $W_p^l(V)$  be the Banach space of complex functions whose derivatives of order  $\leq l$  are  $p$ -summable over  $V$ . One of the imbedding theorems of Sobolev [10] says that the topological inclusion

$$(2.1) \quad W_p^l(V) \subset W_q^m(V)$$

holds, provided, for example, that

$$0 < \frac{n}{p} - l < \frac{n}{q} - m$$

and that  $S$  can be reached by the vertex of a cone of fixed shape contained in  $V$ . We say that  $y \in S$  is regular if it belongs to the boundary of an open part of  $V$  which has the property (2.1). Are there better geometrical criteria than the cone property? Are there necessary and sufficient conditions for regularity resembling the Wiener criterion for regularity in Dirichlet's problem? The results of Campanato [11] are a beginning.

The last question leads us back to elliptic boundary problems. The form given to them previously is very comprehensive, but this is bought with the assumption of smoothness of the boundary. In simple cases, all smoothness can be thrown away. Let  $H^l(V)$  be the closure in  $W_2^l(V)$  of smooth functions with compact supports. If  $V$  is bounded, Dirichlet's problem for a real elliptic  $a(x, D)$  of

order  $2l$  can be put in the form

$$au = 0, \quad u - w \in H^l(V)$$

where  $w \in W_2^l(V)$  is given. Not much is known about how the smoothness of  $u$  in  $V + S$  is influenced by that of  $S$  and  $w$ , but a start has been made by Kondratev [12], who works with a general notion of capacity.

### 3. HYPERBOLIC EQUATIONS

For simplicity, we shall now assume that  $X = R^n$ . A distribution  $E(x, y)$  such that

$$a(x, D) \int E(x, y) f(y) dy = f(x)$$

for smooth functions  $f$  with compact supports is said to be an elementary solution of  $a$  with pole at  $y$ . To begin with, let  $a = a(D)$  have constant coefficients. We say that  $a$  is hyperbolic if it has an elementary solution with support in a cone issuing from the pole, which is proper in the sense that it is contained in some half-space. Let  $(x - y)N \geq 0$  be one of them. It can be shown (see [3]) that an equivalent algebraic condition is the following:  $g(N) \neq 0$  and the zeros of

$$a(p + \tau N) = 0$$

have bounded imaginary parts when  $p$  is real. Observe that the pairs  $a, \pm N$  have this property at the same time. In particular, the zeros of

$$(3.1) \quad g(p + \tau N) = g(N) \prod_1^m [\tau + \lambda_k(p)]$$

are real for  $p$  real; if they are different for  $p \not\equiv 0 \pmod{N}$ , which we shall assume, the inverse implication holds and  $a$  is said to be strongly hyperbolic (with respect to  $N$ ).

*Example 1.* The wave operator  $c^{-2}D_1^2 - D_2^2 - \dots - D_n^2$ ,  $xN = x_1$ ,  $\lambda_{1,2} = c^{-1}p_1 \pm \sqrt{p_2^2 + \dots + p_n^2}$ , ( $c > 0$ ). The characteristic conoid  $\Gamma$  of  $a$  is the real algebraic surface  $g(p) = 0$ . This surface consists of  $m$  different sheets  $\Gamma_1, \dots, \Gamma_m$  given by  $\lambda_1(p) = 0, \dots, \lambda_m(p) = 0$ . Every straight line with direction  $N$  not passing through the origin meets  $\Gamma$  in  $m$  real separate points. The

complement of  $\Gamma$  has  $m + 1$  components (if  $n > 2$ ). The component containing  $N$  is called the characteristic cone  $\dot{\Gamma} = \dot{\Gamma}(g, N)$  of  $a$ . It is convex, and  $-N$  determines the opposite cone.

*Example 2.* In the example above,  $\dot{\Gamma}$  is the cone  $c^{-1}p_1 - \sqrt{p_2^2 + \cdots + p_n^2} > 0$ . The conoid  $\Gamma$  has a dual object, the bicharacteristic conoid  $C$  in  $x$ -space, given in parametric form by

$$x = \pm g_p(p), \quad p \in \Gamma$$

If  $p$  runs through a sheet of  $\Gamma$ ,  $x$  runs through a corresponding sheet of  $C$ . The common part of  $C$  and  $xN \geq 0$  is contained in and bounds the bicharacteristic cone

$$\dot{C} = \text{all } x \text{ with } x \dot{\Gamma} \geq 0$$

*Example 3.* In the foregoing example,  $C$  is the conoid

$$c^2 x_1^2 - x_2^2 - \cdots - x_n^2 = 0$$

and  $C$  the cone

$$cx_1 - \sqrt{x_2^2 + \cdots + x_n^2} \geq 0$$

In the general case,  $a(D)$  has an elementary solution  $E(x - y)$  such that

$$\text{support } E \subset \dot{C}$$

The singularities of  $E$  are on the bicharacteristic conoid  $C$ . Outside  $C$ ,  $E$  is regular and, in fact, analytic.

When  $a$  has variable coefficients and is strongly hyperbolic, we have characteristic conoids  $\Gamma(y)$  and cones  $\dot{\Gamma}(y)$  in  $p$ -space determined by  $g(y, p)$  and a continuous choice of  $0 \neq N(y) \in \dot{\Gamma}(y)$ . The bicharacteristic conoid  $C(y)$  with vertex at  $y$  is no longer generated by straight lines. It is the projection on  $x$ -space of all real bicharacteristic bands

$$dx = g_p(x, p) dt, \quad dp = -g_x(x, p) dt$$

$$x[0, y] = y, \quad p[0, q] = q$$

with  $g(y, q) = 0$ . Restricted locally to a half-space through  $y$ , it bounds the bicharacteristic cone  $\dot{C}(y)$ , whose tangent cone at  $y$  is

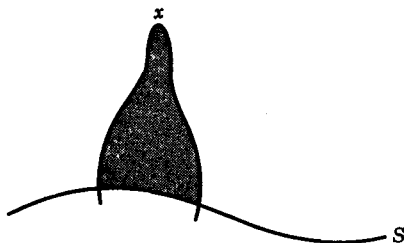
characterized by  $(x - y)\dot{\Gamma}(y) \geq 0$ . The operator  $a(x, D)$  has an elementary solution  $E(x, y)$  with pole at  $y$  and support in  $\dot{C}(y)$  regular outside  $C(y)$  (see [3]), Cauchy's problem

$$au = \ddot{v}, \quad u = w \text{ } m \text{ times on } S$$

(meaning that  $u - w$  vanishes on  $S$  together with all derivatives of order  $< m$ ) has a unique solution for surfaces  $S: s(x) = 0$  which are space-like in the sense that

$$\text{grad } s(x) \in \pm \dot{\Gamma}(x)$$

for all  $x \in S$  (see, e.g., [3]). The value of the solution  $u$  at  $x$  depends only on  $v$  and  $w$  in the compact region common to one side of  $S$  and one of the bicharacteristic cones issuing from  $y$ .



*Progressive Waves.* The two-dimensional wave equation

$$c^2 u_{xx} - u_{tt} = 0$$

has the solution  $u = h(x \pm ct)$ ,  $h$  arbitrary, called progressive waves. Considered as functions of  $x$ , these solutions move (are translated) in opposite directions with velocity  $c$  when the time  $t$  changes. Generalized to a nonlinear situation, they occupy a prominent place in Riemann's work [14] on wave propagation.

Much of the recent progress in the theory of hyperbolic operators has been achieved by the construction of approximate progressive waves in the general case (see, e.g., [13, 15, and 16]). We shall now outline this construction. Let  $\sigma(x)$  be a real regular function with  $\sigma_x = \text{grad } \sigma \neq 0$  and let  $h$  be a distribution in one variable with successive derivatives and integrals

$$h_0 = h, \quad h'_{k+1} = h_k; \quad k = 0, \pm 1, \pm 2, \dots$$

We say that a distribution  $v$  is regular of order  $\geq k^*$  with respect to

the distribution

$$(h \circ \sigma)(x) = h(\sigma(x))$$

if there exist regular  $v_k, v_{k+1}, \dots$  such that the differentiability of

$$v(x) - \sum_k^l v_j(x) h_j \circ \sigma(x)$$

tends to infinity with  $l$ . Notice that

$$a\left(x, \frac{\partial}{\partial x}\right)(F(x)h_0 \circ \sigma(x)) = \sum_0^m \left[ a_k\left(x, \frac{\partial}{\partial x}\right) F(x) \right] h_{k-m} \circ \sigma(x)$$

where

$$\text{order } a_k\left(x, \frac{\partial}{\partial x}\right) \leq k$$

In particular, if

$$a\left(x, \frac{\partial}{\partial x}\right) = g\left(x, \frac{\partial}{\partial x}\right) + g'\left(x, \frac{\partial}{\partial x}\right) + \dots$$

with  $g$  and  $g'$  homogeneous of orders  $m$  and  $m - 1$  respectively, then

$$a_0 = g(x, \sigma_x)$$

$$a_1 = g_p(x, \sigma_x) \frac{\partial}{\partial x} + \frac{1}{2} g_{pp}(x, \sigma_x) \sigma_{xx} + g'(x, \sigma_x)$$

with

$$g_p(x, p) = \text{grad}_p g(x, p), \quad g_{pp} = \frac{\partial^2 g}{\partial p_j \partial p_k}$$

and natural summation conventions. Hence if

$$(3.2) \quad g(x, \sigma_x) = 0, \quad a_1\left(x, \frac{\partial}{\partial x}\right) F(x) = 0$$

and

$$(3.3) \quad u(x) = F(x)h \circ \sigma(x)$$

then  $au$  is regular of order  $\geq 2 - m$ , that is, two units more than expected. Observe that the equations (3.2) do not involve  $h$  and that they have plenty of solutions. More generally, it is easy to

see that we can choose

$$(3.4) \quad u(x) = \sum_0^l F_j(x) h_j \circ \sigma(x)$$

so that  $au$  is regular of order  $\geq l + 2 - m$ . In the analytic case we can even let  $l \rightarrow \infty$  and obtain solutions of  $au = 0$  depending among other things on an arbitrary distribution  $h$ .

It is clearly reasonable to consider (3.3) and (3.4) as approximate progressive waves with the wave fronts

$$\sigma(x) = \text{constant}$$

Notice that if  $\lambda_k(x, p)$  is defined by factoring  $g(x, p + \tau N)$  [see (3.1)], then (3.2) means that

$$\lambda_k(x, \sigma_x) = 0, \quad a_1 \left( x, \frac{\partial}{\partial x} \right) F(x) = 0$$

for one  $k$  and that  $\sigma$  and  $F$  can be described on any space-like surface  $S$ . In particular, there are  $m$  different wave fronts reducing to a given one in  $S$ . An elementary solution having a pole at  $y \in S$  and vanishing on one side of  $S$  is an example of this solution: it can be looked upon as a sum of  $m$  progressive waves starting from a point in  $S$ . Approximate progressive waves can be used to give existence proofs for elementary solutions and for Cauchy's problem. So far, proofs of this kind require more smoothness of the coefficients of  $a$  than the proofs using energy methods, but they give more information about the irregularities.

Most physical boundary problems for hyperbolic equations are of mixed type in the sense that they involve reflecting barriers. A lot of work has been done with these problems (see, e.g., [17] and [18] and the literature quoted there), but no systematic theory is in sight. It should be noted that the theory of reflection of approximate progressive waves is an interesting but at present neglected field of study.

Wave propagation also has an ergodic aspect: when there is a complicated enough mixture of waves, the theory has to be restricted to asymptotic distributions of energy for large times. In some simple cases there is equidistribution in space for the high frequencies (see [19]); recently the energy distribution outside a reflecting obstacle has been analyzed by Lax, Morawetz, and Phillips [20] with very interesting results.

## 4. FROBENIUS' THEOREM

Let  $M$  be a  $C^\infty$ -module of homogeneous real first-order linear differential operators

$$A = \sum_1^n a_j(x) \frac{\partial}{\partial x_j}$$

which has the property that

$$A, B \in M \Rightarrow [A, B] = AB - BA \in M$$

and let  $M'$  be all distributions  $f = f(x)$  for which  $Af = 0$  for every  $A$  in  $M$ . Let  $r(x)$  be the linear dimension of  $M$  at the point  $x$ . Close to a point  $x$  where  $r$  is constant we have the classical theorem of Frobenius [21]: there exist coordinates  $y_1, \dots, y_n$  such that  $M$  is spanned by  $\partial/\partial y_1, \dots, \partial/\partial y_r$ . In particular,  $M'$  is the set of distributions depending only on  $y_{r+1}, \dots, y_n$ .

The assumption that  $r(x)$  is constant is violated in many interesting cases, for example, when the elements of  $M$  are the infinitesimal generators of a connected Lie group  $G$  acting on  $C^\infty(X)$ . In that case,  $M'$  is the set of distributions invariant under  $G$ . When  $G$  is the Lorentz group and  $X$  is a product of  $g$  copies of  $R^4$  where  $g$  is the dimension of  $G$ ,  $M'$  has some interest in connection with quantum field theory. It is known for  $g = 1$  ([22]), and has been investigated for  $g = 2$  ([23]). But these are only partial results, and we would like to have an answer to the following question. What is the local structure of  $M$  and  $M'$  in a neighborhood of a point  $x$  where  $r$  is discontinuous? If the coefficients of all  $A \in M$  are real analytic, a complete answer is perhaps possible.

*Lecture II    Uniformization in Cauchy's Problem*

## 5. INTRODUCTION

Let  $X$  be a complex analytic nonsingular variety of dimension  $l$  with coordinates  $x_1, \dots, x_l$ , and let

$$a\left(x, \frac{\partial}{\partial x}\right) = \sum_{|\alpha| \leq m} a_\alpha(x) \left(\frac{\partial}{\partial x}\right)^\alpha$$

be a linear differential operator on  $X$  of order  $m$  with holomorphic (regular analytic) coefficients. Let

$$g\left(x, \frac{\partial}{\partial x}\right) = \sum_{|\alpha|=m} a_{\alpha}(x) \left(\frac{\partial}{\partial x}\right)^{\alpha}$$

be its principal part. Let  $S: s(x) = 0$  be a nonsingular analytical hypersurface in  $X$  and consider Cauchy's problem with holomorphic data  $v$  and  $w$ :

$$(5.1) \quad \begin{aligned} a\left(x, \frac{\partial}{\partial x}\right) u(x) &= v(x) \text{ in } X \\ u(x) - w(x) &= 0 \quad m \text{ times on } S \end{aligned}$$

According to theorem of Cauchy-Kowalewski, this problem has a unique holomorphic solution near a point  $x \in S$  which is non-characteristic in the sense that

$$g(x, s_x) \neq 0$$

where  $s_x = \text{grad } s$ . In this connection we may ask ourselves the following question: what happens to the solution near the characteristic point of  $S$ ? Take for instance

$$a\left(x, \frac{\partial}{\partial x}\right) = \frac{\partial}{\partial x_1}$$

$$S: x_2 - x_1^p = 0$$

Then

$$T: x_1 - x_2 = 0$$

is a characteristic subvariety of  $S$  of codimension 1. It is part of a characteristic hyperplane

$$K: x_2 = 0$$

which touches  $S$  along  $T$ . More precisely, the order of contact is  $p$ . The solution of Cauchy's problem is

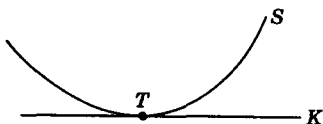
$$u(x) = w(y_1, x_2, \dots) + \int_{y_1}^{x_1} v(s, x_2, \dots) ds, \quad y_1 = \sqrt[p]{x_2}$$

It is not holomorphic, but it is ramified  $p$  times around  $K$ . The

substitution

$$x_1, x_2, \dots \rightarrow x_1, t^p, \dots$$

uniformizes  $u(x)$ .



In [4], J. Leray proved that this situation is general: the solution is ramified around a characteristic subvariety tangent to the initial surface, and it can be uniformized explicitly.

In the following, which is a report on a long paper by Gårding, Leray, and Kotake [2], all this will be proved in a simple fashion. Actually, [2] deals with systems, but we shall restrict ourselves to a single operator. Following [2] we shall introduce an extra variable  $\xi$  in (5.1), replacing  $S$  by  $S(\xi): s(x) = \xi$ . The problem is then

$$(5.2) \quad a\left(x, \frac{\partial}{\partial x}\right) u(\xi, x) = v(\xi, x) \text{ in } X$$

$$u(\xi, x) - w(\xi, x) = 0 \quad m \text{ times on } S(\xi)$$

This problem is a special case of (5.1) in  $(l+1)$  variables, and the Cauchy-Kowalewski theorem still holds. The uniformization of (5.2) is very simple. Let  $\xi(t, x)$  be the solution of Hamilton-Jacobi's equation

$$\xi_t + g(x, \xi_x) = 0, \quad \xi(0, x) = s(x)$$

We shall see that the substitution

$$\xi, x \rightarrow \xi(t, x), x$$

uniformizes the solution of (5.2). Its restriction to  $\xi(t, x) = 0$  uniformizes the solution of (5.1).

## 6. AN INVARIANT

Let  $p = (p_1, \dots, p_l)$  denote a covector transforming like a gradient. A straightforward computation gives

$$e^{\lambda f(x)} a\left(x, \frac{\partial}{\partial x}\right) e^{\lambda f(x)} = g(x, f_x) \lambda^m + [\tfrac{1}{2} g_{pp}(x, f_x) f_{xx} + g'(x, f_x)] \lambda^{m+1} + \dots$$

where  $a(x, \partial/\partial x) = g(x, \partial/\partial x) + g'(x, \partial/\partial x) + \dots$  with  $g$  and  $g'$  homogeneous of order  $m$  and  $m - 1$  and

$$g_{pp}(x, f_x)f_{xx} = \sum \frac{\partial^2 g(x, f_x)}{\partial p_j \partial p_k} \frac{\partial^2 f}{\partial x_j \partial x_k}$$

Hence,  $g(x, p)$  is invariant under coordinate transformations and so is

$$\frac{1}{2}g_{pp}(x, f_x)f_{xx} + g'(x, f_x)$$

## 7. A LEMMA

The main result, which follows in the next section, will be proved by a reduction to a special case which we will treat now. Let  $t$  be an additional variable, let

$$b_{jk}t, \frac{x\partial}{\partial x}; \quad j, k = 1, \dots, m$$

be holomorphic differential operators, and consider the Cauchy problem

$$(7.1) \quad \frac{\partial u_j}{\partial t} = \sum b_{jk}u_k + v_j$$

$$u_j(0, t) = w_j(0, x)$$

Furthermore, suppose that there exist integers  $m_k$  such that

$$(7.2) \quad \text{order } b_{jk} \leq m_j - m_k + 1$$

It is no restriction to assume that all  $m_j \geq 0$ .

*Lemma 1.* If the  $b_{jk}$ ,  $v_j$  and  $w_j$  are holomorphic for

$$|t| \leq r, |x| = |x_1| + \dots + |x_1| \leq R$$

then (7.1) has a unique solution which is holomorphic for  $|x| < R$  and  $t$  small enough,  $|t| < \text{const}(R - |x|)$ .

REMARK 1. The uniqueness is obvious since (7.1) permits a calculation of the power series for  $u$  around the origin.

REMARK 2. The proof uses the following lemma.

*Lemma 2.* (Rosenbloom [5], Hörmander [3]) If  $u(x)$  is holomorphic for  $|x| < R$  and

$$|u(x)| \leq \frac{r}{(R - |x|)^r}$$

then

$$\left| \frac{\partial u}{\partial x_k} \right| \leq \frac{3(r+1)}{(R-|x|)^{r+1}}$$

PROOF. It suffices to take  $l = 1$ . Then

$$\frac{\partial u}{\partial x} = \frac{1}{2i} \int_{\gamma} \frac{u(y)}{(x-y)^2} dy$$

where  $\gamma$  is the circle  $|x-y| = \epsilon$ ;  $0 < \epsilon < R - |x|$ . Hence

$$\left| \frac{\partial u}{\partial x} \right| \leq \frac{r}{\epsilon(R-|x|-\epsilon)^r}$$

Putting  $\epsilon = (R - |x|)/(r+1)$  we get the desired result since

$$\left( \frac{r+1}{r} \right)^r < e < 3$$

PROOF OF LEMMA 1. We use successive approximation as follows

$$\frac{\partial u_{j,0}}{\partial t} = v_j, \quad u_{j,0}(0, x) = w_j(x)$$

$$\frac{\partial u_{j,r}}{\partial t} = \sum b_{jk} u_{k,r-1}, \quad u_{k,r}(0, x) = 0$$

for  $r > 0$ . Then the series

$$(7.3) \quad u_j(t, x) = \sum_0^{\infty} u_{j,r}(t, x)$$

converges to a solution. In fact, suppose that we have the estimate

$$(7.4) \quad |u_{k,r-1}| \leq cC^{r-1}r^{m_k}|t|^{r-1}/(R-|x|)^{r+m_k}$$

which holds for  $r = 1$  provided  $c$  is large enough. Then

$$\sum_k |b_{jk} u_{k,r-1}| \leq cC^r r^{1+m_j} |t|^{r-1} / (R-|x|)^{r+1+m_j}$$

provided  $C$  is large enough, depending on  $b_{jk}$ ,  $m$  and  $R$ . Integrating we find (7.4) for  $r$  and the series (7.3) converges for  $C|t| < R - |x|$ .

# 8. UNIFORMIZATION

We shall use a change of variables

$$(8.1) \quad (t, x) \rightarrow (\xi(t, x), x)$$

$\xi$  being holomorphic. Let us put

$$(u \circ \xi)(t, x) = u(\xi(t, x), x)$$

so that, with indices denoting derivatives, we have

$$(8.2) \quad \frac{\partial}{\partial t} u \circ \xi = \xi_t u_\xi \circ \xi, \quad \frac{\partial}{\partial x} u \circ \xi = \xi_x u_\xi \circ \xi + u_x \circ \xi$$

We say that a function

$$u(\xi, x)$$

is uniformized of order  $n$  by (8.1) if  $v \circ \xi$  is holomorphic in  $t, x$  for any derivative  $v$  of  $u$  of order  $\leq n$ . In view of (8.2), it suffices for this that  $v$  is holomorphic when  $v$  is one of the functions

$$u_j(\xi, x) = \left( -\frac{\partial}{\partial \xi} \right)^j u(\xi, x)$$

and  $0 \leq j \leq n$ . The next lemma is obtained from (8.2) by a straightforward computation.

*Lemma 3.* Let

$$a(x, \partial/\partial x) = g(x, \partial/\partial x) + g'(x, \partial/\partial x) + \dots$$

with  $g$  and  $g'$  homogeneous of order  $m$  and  $m-1$ . Then

$$\left[ a \left( x, \frac{\partial}{\partial x} \right) u \right] \circ \xi = \sum_{j=0}^m a_j \left( x, \xi_x, \xi_{xx}, \dots, \frac{\partial}{\partial x} \right) (u_{m-j} \circ \xi)$$

where order  $a_j \leq j$  and, explicitly,

$$a_0 = g(x, \xi_x)$$

$$a_1 = g_p(x, \xi_x) \frac{\partial}{\partial x} + \frac{1}{2} g_{pp}(x, \xi_x) \xi_{xx} + g'(x, \xi_x)$$

REMARK. The  $a_j$  are independent of the coordinate system. We have seen in Section 2 that

$$h(t, x) = \frac{1}{2} g_{pp}(x, \xi_x) \xi_{xx} + g'(x, \xi_x)$$

is also invariant.

Let us now pass to the variables  $t, x$  in (5.2). The result is

$$(8.3) \quad \sum a_j \left( x, \xi_x, \frac{\partial}{\partial x} \right) (u_{m-j} \circ \xi) = v \circ \xi$$

$$u_j \circ \xi = w_j \circ \xi \text{ for } t = 0, \quad 0 \leq j < m$$

which can be written as follows

$$(8.4) \quad \frac{\partial}{\partial t} u_j \circ \xi = -\xi_t u_{j+1} \circ \xi, \quad 0 \leq j < m-1$$

$$g(x, \xi_x) \frac{\partial}{\partial t} u_{m-1} \circ \xi - \xi_t \sum_1^m a_j \left( x, \xi_x, \frac{\partial}{\partial x} \right) u_{m-j} \circ \xi = -\xi_t v$$

$$u_j \circ \xi = w_j \circ \xi \text{ for } t = 0, \quad 0 \leq j < m$$

Now choose  $\xi(t, x)$  so that it satisfies the Hamilton-Jacobi equation

$$(8.5) \quad \xi_t + g(x, \xi_x) = 0$$

$$\xi(0, x) = s(x)$$

Notice that  $\xi(t, x) = 0$  is a regular hypersurface in  $t, x$ -space for  $t$  small. Then, at the noncharacteristic points of  $S(\xi)$  where  $g(x, \xi_x) \neq 0$  (8.5) is equivalent to

$$(8.6) \quad \frac{\partial}{\partial t} u_j \circ \xi + \xi_t u_{j+1} \circ \xi = 0, \quad 0 \leq j < m-1$$

$$\frac{\partial}{\partial t} u_{m-1} \circ \xi + \sum_1^m a_j \left( x, \xi_x, \frac{\partial}{\partial x} \right) u_{m-j} \circ \xi = v \circ \xi$$

$$u_j \circ \xi = w_j \circ \xi \text{ for } t = 0, \quad 0 \leq j < m$$

which, by Lemma 1, has a holomorphic solution (choose  $m_j = j$ ). Hence, by analytical continuation we have a uniformization theory.

*Uniformization Theorem.* If

$$g(x, \xi_x) \neq 0 \text{ on } S(\xi)$$

and if  $\xi$  solves (8.5), then Cauchy's problem (5.2) has a solution  $u(\xi, x)$  which, close to  $S(\xi)$ , is uniformized of order  $< m$  by the substitution  $u \rightarrow u \circ \xi$ : Its restriction to  $\xi(t, x) = 0$  uniformizes  $u(0, x)$  of order  $< m$ .

## 9. GEOMETRY OF THE UNIFORMIZATION

The uniformizing mapping

$$(9.1) \quad t, x \rightarrow \xi, x = \xi(t, x), x$$

sends  $t = 0$  into  $S(\xi)$  and its Jacobean is  $\xi_t$ . A function  $u(\xi, x)$  which is uniformized by (9.1) is in general multivalued and ramified around the image  $K(\xi)$  in  $\xi, x$ -space of the variety  $\xi_t = 0$ . We shall give a description of  $K(\xi)$ .

The Hamilton-Jacobi equation is closely connected with the Hamiltonian system

$$(9.2) \quad \begin{aligned} dx &= g_p(x, p) dt, & dp &= -g_x(x, p) dt \\ x[0, y] &= y, & p[0, y] &= s_y(y) \end{aligned}$$

The change of variables  $(x, t) \rightarrow (y, t)$  is holomorphic both ways for small  $t$  and reduces to the identity of  $t = 0$ . We shall express it by writing

$$(9.3) \quad f(t, x) = f[t, y]$$

$f(\quad)$  and  $f[\quad]$  denoting different functions.

Since  $dg(x, p) = (g_x g_p - g_p g_x) dt = 0$ ,  $g(x, p)$  is invariant on the curve (band) (9.2) and since  $p dx = mg(x, p) dt$ , the band is characteristic if  $g(y, s_y) = 0$ , that is, if it starts from a characteristic contact element  $y, s_y(y)$ .

Furthermore, the form

$$\omega = p dx - g dt$$

is a relative invariant for (9.2). In fact, its differential depends only on the initial contact element (Cartan [1]). In our case,  $\omega = ds(y)$  for  $t = 0$ , so that  $\omega$  is exact. Hence, there exists a func-

tion  $\xi[t, y]$  such that

$$d\xi[t, y] = \omega, \quad \xi[0, y] = s(y)$$

Changing to the variables  $t, x$ , putting

$$\xi(t, x) = \xi[t, y]$$

we get

$$\begin{aligned} \xi_x &= p[t, y], & \xi_t &= -g(x, p) \rightarrow \xi_t + g(x, \xi_x) = 0 \\ \xi(0, x) &= s(x) \end{aligned}$$

so that  $\xi$  is the unique solution of (8.5). Integrating for  $y$  constant we also get

$$d\xi[t, y] = (m-1)g(x, p) dt = (m-1)g(y, s_y) dt$$

so that, more explicitly,

$$\xi(t, x) = s(y) + (m-1)tg(y, s_y)$$

If we now define a bicharacteristic curve as the projection of (9.2) into  $X$ -space:  $x = x[t, y]$  issuing from a characteristic contact element of  $S(\xi)$ , we get the following lemma:

*Lemma 4.*  $K(\xi)$  is the union of bicharacteristics issuing from the characteristic part  $T(\xi)$  of  $S(\xi)$ . Consider the differential operator

$$L = \frac{\partial}{\partial t} + a_1$$

which, if

$$\frac{d}{dt} = \frac{\partial}{\partial t} + g_p(x, \xi_x) \frac{\partial}{\partial x}$$

denotes differentiation along the curve  $x = x[t, y]$  (i.e., for  $y$  constant), can be written as

$$L = \frac{d}{dt} + h(t, x)$$

Since  $dg/dt = 0$ ,  $L$  commutes with  $g$ . For future use we note the following:

Lemma 5. We have

$$(9.4) \quad g(u_m \circ \xi - w_m \circ \xi) = (v - aw) \circ \xi + t \operatorname{hol}(t, x)$$

$$(9.5) \quad Lgu_m \circ \xi = g \operatorname{hol}(t, x)$$

PROOF. By (8.2)

$$gu_m \circ \xi = \frac{\partial}{\partial t} u_{m-1} \circ \xi = \operatorname{hol}(t, x)$$

and, applying Lemma 3 to  $a(u - w) = v - aw$ , we have

$$g(u_m - aw_m) \circ \xi + \sum_1^m a_j(u_{m-j} - w_{m-j}) \circ \xi = (v - aw) \circ \xi$$

Since the  $a_1, a_2, \dots$  do not contain  $\partial/\partial t$ , (8.4) shows that the sum on the left vanishes with  $t$ . Hence (9.4) follows. Applying Lemma 3 to  $au_1 = v_1$ , we have (8.4) with  $m$  replaced by  $m+1$ . Hence, since  $g$  commutes with  $L$ , we have

$$Lgu_m \circ \xi + g \sum_2^m a_j u_{m+1-j} \circ \xi = gv_1 \circ \xi$$

so that (9.5) follows.

## 10. THE HIGHER DERIVATIVES

Since

$$u_{j+1} \circ \xi = \frac{1}{g} \frac{\partial}{\partial t} u_j \circ \xi, \quad j \geq 0$$

$$u_j \circ \xi = \operatorname{hol}(t, x), \quad 0 \leq j < m$$

we have

$$u_{m+j-1} \circ \xi = f_0 + \frac{f_1}{g} + \dots + \frac{f_j}{g^j}, \quad j > 0$$

where  $f_0, \dots, f_j$  are holomorphic in  $t, x$ , but not uniquely determined. Clearly  $f_2$  is unique on  $g=0$ , a choice of  $f_2$  determines  $f_3$  on  $g=0$  and so on. It is possible to give explicit formulas for these functions and we shall do this for  $j=1$ . Let us write

$$F(t, x) \equiv 0 \bmod G(t, x)$$

when

$$F(t, x) = G(t, x) \text{ hol } (t, x)$$

*Theorem 1.* Let  $U(t, x)$  be the solution of

$$\left[ \frac{d}{dt} + h(t, x) \right] U(t, x) = 0,$$

$$U(0, x) = (v - aw) \circ \xi(0, x)$$

Then,

$$(10.1) \quad u_m \circ \xi = w_m \circ \xi + \frac{U(t, x)}{g(x, \xi_x)} \text{ mod } t$$

REMARK. Let  $G(t, x)$  be the solution of

$$\left[ \frac{d}{dt} + h(t, x) \right] G(t, x) = 0$$

$$G(0, x) = 1$$

along the characteristic band (9.2). Then

$$U(t, x) = G(t, x)[v(\xi, y) - a\left(y, \frac{\partial}{\partial y}\right)w(\xi, y)]_{\xi=s(y)}$$

PROOF. Put

$$V(t, x) = g(u_m \circ \xi - w_m \circ \xi) - U(t, x)$$

By Lemma 5,

$$LV \equiv 0 \text{ mod } g$$

$$V \equiv 0 \text{ mod } t$$

so that, passing to the variables  $t, y$

$$\left( \frac{d}{dt} + h[t, y] \right) V[t, y] = g(y, s_y) \text{ hol } [t, y]$$

$$V[0, y] = 0$$

It follows from this that

$$V[t, y] = tg(y, s_y) \text{ hol } [t, y]$$

that is,

$$V \equiv 0 \text{ mod } tg$$

which is equivalent to (10.1). The proof of the remark is obvious.

## REFERENCES FOR LECTURE I

1. P. C. Rosenbloom, "Linear partial differential equations," *Surveys Appl. Math.* **5** (1958) 43-195.
2. L. Gårding, "Some trends and problems in the theory of linear partial differential operators," *Proc. Int. Congress of Math.*, Edinburgh (1958).
3. L. Hörmander, *Partial differential operators*, Academic Press, New York, Berlin (1963).
4. J. Peetre, "Another approach to elliptic boundary problems," *Comm. Pure Appl. Math.* **14** (1961) 711-731.
5. J. B. Lopatinski, "On a method of reducing boundary problems for a system of elliptic type to regular equations," *Ukrain. Mat. Zur.* **5** (1953) 123-151.
6. F. Noether, "Über eine Klasse singulärer Integralgleichungen," *Math. Ann.* **82** (1921) 42-63.
7. F. Atiyah and I. M. Singer, "The index of elliptic operators on compact manifolds," to be published.
8. R. Bott, "The index theorem for homogeneous differential operators," to be published.
9. J. Peetre, "Mixed problems for higher order elliptic equations in two variables I," *Annali della Scuola Normale Superiore di Pisa. Serie III, Vol. XV, Fasc. IV* (1961) 337-353.
10. S. L. Sobolev, "Some applications of functional analysis to mathematical physics," *Izdat. Leningrad. Gos. Univ.* (1950).
11. S. Campanato, "Il teorema di immersione di Sobolev per una classe di aperti non dotati della proprietà di cono," *Ricerche Mat.* **11** (1962) 103-122.
12. V. A. Kondratev, "On the solvability of the first boundary value problem for elliptic equations," *Dokl. Akad. Nauk SSSR* **36** (1960), 771-774.
13. R. Courant and P. Lax, "The propagation of discontinuities in wave motion," *Proc. Nat. Acad. Sci.* **42** (1956) 872-876.
14. B. Riemann, *Ueber die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite*. Collected Works, Dover Publications.
15. P. Lax, "Asymptotic solutions of oscillatory initial value problems," *Duke Math. J.* **24.4** (1957) 627-646.
16. G. Ludwig, "Exact and asymptotic solutions of the Cauchy problem," *Pure and Appl. Math.* **13.3** (1960) 473-508.
17. V. Thomée, "Existence proofs for mixed problems in the theory of linear hyperbolic equations in two variables by means of the continuity method," *Math. Scand.* **6** (1958) 5-32.
18. C. F. D. Duff, "Mixed problems for hyperbolic equations of general order," *Canad. J. Math.* **9** (1959) 195-221.
19. T. Carleman, "Propriétés asymptotiques des fonctions fondamentales des membranes vibrantes." *C. R. du huitième Congrès des mathématiciens scandinaves*, Stockholm (1934).
20. P. D. Lax, C. S. Morawetz, and R. S. Phillips, "The exponential decay of solutions of the wave equation in the exterior of a star-shaped obstacle," *Bull. Amer. Math. Soc.* **68.6** (1962) 593-595.
21. G. Frobenius, "Über das Pfaßsche problem," *J. Math.* **82** (1877) 230-315.

22. P. Methée, "Sur les distributions invariantes dans le groupe des rotations de Lorentz," *Comm. Math. Helv.* **28** (1954) 224-261.
23. L. Gårding, "Distributions invariantes," Séminaire sur les équations aux dérivées partielles. Collège de France. November 1961-May 1962 (mimeographed).

## REFERENCES FOR LECTURE II

1. E. Cartan, *Les invariants intégraux*, Paris (1932).
2. L. Gårding, T. Kotake, and J. Leray, "Uniformisation et développement asymptotique de la solution du problème de Cauchy linéaire, à données holomorphes (Problème de Cauchy I<sup>bis</sup> et VI.)" To be published in *Journal de Mathématiques, pures et appliquées*.
3. L. Hörmander, *Partial differential operators*, Berlin (1963).
4. J. Leray, "Uniformisation de la solution du problème analytique de Cauchy, près de la variété qui porte les données de Cauchy (Problème de Cauchy I.)" *Bull. Soc. Math. France* **6.85** (1957) 389-429.
5. P. C. Rosenbloom, The Majorant method, *Proc. Sympos. Pure Math.* IV 51-72. Am. Math. Soc., Providence, R.I. (1961).

# 5

## Quasiconformal Mappings and Their Applications

*Lars V. Ahlfors*

### 1. INTRODUCTION

It would be hopeless and confusing to try to trace the modern development in the whole wide field of analysis. For instance, the recent years have seen an enormous upswing in the theory of partial differential equations, a theory that heretofore had been limited to the comparatively mildmannered equations that occur in physical applications. On another front, the theory of analytic functions of several variables and of manifolds with complex structures have received a strong impetus from methods of topology and algebraic geometry. Since the interactions between various branches of mathematics have a healthy tendency to grow stronger rather than weaker, such progress cannot be ignored even in a partial survey, but this chapter shall be restricted to its indirect repercussions. The core of classical analysis is the theory of functions of one complex variable with its direct traditions from Cauchy and Riemann. In some respects this is a very narrow area, but in contrast with other branches of analysis it is one that was already highly developed at the turn of the century, and in which further progress is therefore both difficult and very significant.

Let us first throw a quick glance at the status of classical function theory in the 1920's, say, a period that set the stage for this author's generation. The general theory of uniformization had just been completed, and thereby a formidable stumbling block had been removed. R. Nevanlinna had founded his monumental theory of meromorphic functions which by its very perfection seemed to block further progress in this direction. Many mathematicians were intrigued by the results achieved in the theory of univalent functions, and whether we believed in its profundity or dismissed it as a curiosity, the Bieberbach conjecture was and still remains an unmistakable challenge.

In view of the difficulty and abundance of the problems, it seemed clear that new tools would be needed. Such tools were indeed forthcoming. This author admires the variational method of M. Schiffer, which seemed to bring the Bieberbach conjecture within easy reach. Another essential contribution was the Bergman kernel function which established the first connection with Hilbert space theory. The theory of extremal length which had its roots in the work of H. Grötzsch, but rose from obscurity to brilliance thanks to the vision of A. Beurling should also be mentioned.

Equally important, and in their ultimate consequences perhaps overwhelmingly so, were the observations that poured forth from the uncanny brain of Teichmüller. For better or for worse, his ghost has risen from an early grave to haunt a great deal of the thinking that has gone on in postwar function theory. An excellent way to acquaint the mathematical public with the trends in my special field is to relate some of the experiences of this author and fellow specialists in dealing with the heritage of Teichmüller.

It should be emphasized that this brief survey does not and cannot do justice to all that has been done in the field of one complex variable. To mention all would be to spread it out so thin that nobody would be much wiser. The intention is quite the contrary, namely to disseminate knowledge of a method whose very name might easily create the false impression of a spineless generalization of conformal mapping. Our aim will be achieved if we can convey that quasiconformal mappings are intimately and essentially, even surprisingly, connected with the most classical problems in the theory of analytic functions.

## 2. QUASICONFORMAL AND CONFORMAL MAPPINGS

Historically, the first mention of quasiconformal mappings (by a different name) is in a small paper in 1928 of H. Grötzsch. He solves the extremal problem of mapping a square on a rectangle so that the vertices correspond, and so that the mapping is as nearly conformal as possible. The problem was treated primarily as a curiosity, but the paper also contains a more significant remark to the effect that Picard's theorem is valid for quasiconformal mappings.

Like all of Grötzsch's work, this paper was buried in a small journal and therefore very slow to gain recognition. This author learned of it by word of mouth in 1931, but did not use it until 1936 in a paper that was primarily devoted to quite different questions. Meanwhile, in 1935, Lavrentieff had published a paper on what he called "*fonctions presque-analytiques*" and which were practically the class considered by Grötzsch. It is not clear how the name "quasiconformal mappings" originated, but by this time the notion began to be generally accepted. In the beginning it emerged primarily as a tool in the study of conformal mappings. At the same time it was used, quite independently, by Morrey and the school of Lavrentieff for the study of elliptic equations in two variables. In a few years it became increasingly clear that quasiconformal mappings merited a theory of their own, a view that gained its strongest support from the more prophetic than rigorous writings of Teichmüller in 1938–1940.

Now, to begin the mathematical part of this chapter, our first concern is with the precise definition of a quasiconformal mapping. If  $f$  is a differentiable mapping, it is clear that infinitesimal circles are mapped on infinitesimal ellipses. The ratio  $a/b$  of the major and minor axes of the ellipse is the dilatation at a point, and the mapping is quasiconformal if the dilatation is bounded. Analytically, it is convenient to write

$$df = f_z dz + f_{\bar{z}} d\bar{z}$$

It is sufficient to consider the case where  $f$  is sensepreserving. Then  $|f_{\bar{z}}| < |f_z|$ , and the axes of the ellipse are  $(|f_z| + |f_{\bar{z}}|)|dz|$  and  $(|f_z| - |f_{\bar{z}}|)|dz|$ . The condition of quasiconformality is thus

$$\frac{|f_z| + |f_{\bar{z}}|}{|f_z| - |f_{\bar{z}}|} \leq K < \infty$$

and when  $K$  is given we speak of a  $K$ -quasiconformal mapping. The same condition can be expressed by  $|f_z| \leq k|f_{\bar{z}}|$  with  $k < 1$ . It is obvious from the definition that a mapping and its inverse have the same dilatation, and also that a conformal mapping does not change the dilatation. Similarly, the product of two quasiconformal mappings is again quasiconformal. In other words, the basic composition properties of conformal mappings remain in force.

It was recognized quite early that differentiability was too strong a condition, and in the beginning makeshift arrangements were introduced to allow for a more general behavior at least at some points. But the most important need was for compactness, which can be attained in two ways—geometrically and analytically. Geometrically, a quadrilateral  $Q$  is the conformal image of a rectangle, and the ratio of the sides of the rectangle serves as the modulus  $m$  of  $Q$ . Suppose that the homeomorphism  $f$  maps  $Q$  on  $Q'$  with modulus  $m'$ . Then  $f$  is said to be  $K$ -quasiconformal if  $m'/m \leq K$  for all  $Q$ . It is easy to see that this coincides with the original definition when  $f$  is differentiable, and that the resulting class of  $K$ -quasi conformal mappings is compact.

Analytically, we define a  $K$ -quasiconformal mapping as one which satisfies a partial differential equation

$$(2.1) \qquad f_{\bar{z}} = \mu(z)f_z$$

with  $|\mu| \leq k = (K - 1)/(K + 1)$ . The coefficient  $\mu$  is called the complex dilatation. If it is continuous, we have the same class as before, but the point is that  $\mu$  need only be measurable and  $f$  a weak solution of the equation. More precisely,  $f$  is to be a homeomorphism whose distributional derivatives are locally square integrable and satisfy (2.1).

It is a nontrivial and important fact that these definitions are equivalent. As a sad commentary on the communications between mathematicians, we might mention that the essence of the theorem is contained in Morrey's paper of 1938, but that its relevance for quasiconformal mappings was not clearly realized until 1956 (by Bers).

Equation (2.1) is classical, and is named after Beltrami who wrote it in a different form. Beltrami obtained it as an analogue of the Cauchy-Riemann equations on a surface in space. A little more generally, consider a Riemannian metric in the  $z$ -plane. Its funda-

mental form can be written as

$$ds^2 = \rho(z)^2 |dz + \mu d\bar{z}|^2$$

with  $\rho(z) > 0$ ,  $|\mu| < 1$ . Clearly, a mapping  $w = f(z)$  is conformal from the Riemannian metric to the euclidean metric of the  $w$ -plane, if and only if the ratio  $df/(dz + \mu d\bar{z})$  is independent of the direction of  $dz$ , and this is so exactly when (2.1) is fulfilled. Allowing for a sufficiently general concept of a Riemannian metric, we recognize the fundamental fact that a *quasiconformal mapping is nothing more than a conformal mapping with respect to a metric determined by the complex dilatation*.

This observation, by itself, would be quite sufficient to give quasiconformal mappings a respectable status, but we shall see that the tie-in with analytic functions is much more direct.

### 3. EXTREMAL PROBLEMS

It has already been mentioned that quasiconformal mappings were originally used primarily as a tool in the study of conformal mappings. Many problems in function theory, especially if connected with geometric questions, require the use of cleverly selected auxiliary functions or mappings. If we insist that the mappings be conformal, the construction may be very difficult, because conformal mappings are by nature rather rigid. In many cases quasiconformal mappings will serve the same purpose, and they are much easier to find. For instance, many problems concerned with the parabolic or hyperbolic type of Riemann surfaces were solved by this technique. We shall not dwell on this aspect, because it is relatively uninteresting. Quasiconformal mappings have found their place in the toolkit of the well-equipped analyst, and that is that.

Much more important is the connection with extremal problems. Mathematicians have always been fascinated by maxima and minima, both for their own sake and because of the special circumstances that occur at extreme points. Nowhere are extremal problems as fruitful as in the theory of analytic functions. Analytic proofs depend extensively on estimates, and to find the best possible estimate is not only a source of esthetic satisfaction, but will almost invariably lead to a better understanding of the situation being studied.

As has been pointed out, the very genesis of quasiconformal mappings was connected with the elementary extremal problem formulated by Grötzsch. Teichmüller was the first to extract a general principle: In a class of mappings it is required to find one whose maximal dilatation is a minimum. It is to be expected that the solution is unique, and that the extremal mapping is characterized by simple properties.

Let us illustrate by a few examples. First, as a very slight generalization of Grötzsch's problem we may look for the extremal mapping of one quadrilateral on another. Because dilatation is invariant under conformal mapping it is natural to replace the quadrilaterals by conformally equivalent rectangles. Also, it is to be expected, and quite easy to prove, that the extremal mapping is represented by the affine mapping of one rectangle on the other. Let us now replace the quadrilaterals by pentagons. Teichmüller proves that they can be mapped conformally, on polygons, as shown in Figure 1, such that  $a:b = a':b'$ . The extremal mapping will be represented by a stretching in the direction of the  $y$ -axis. This is already a sophisticated result, mainly because it is not known beforehand which side will contain the concave angle.

The problem is more complicated for hexagons, but can still be solved by the same method. More general problems call for the extremal mapping of a multiply connected region on another, or even of a closed Riemann surface on another of the same genus. Efforts to solve these problems had failed, but in a flash of genius Teichmüller came forth with a daring conjecture which he himself admitted was based on almost nothing. In the few cases that could be solved explicitly he observed that the solution was always given

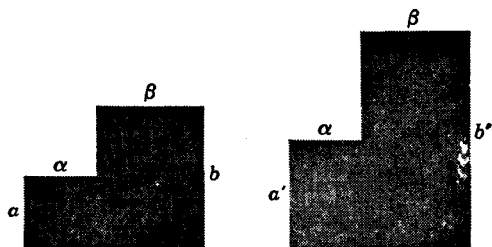


Figure 1

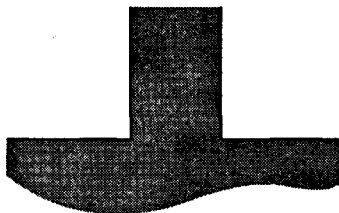
by a conformal mapping, followed by an affine mapping, and by another conformal mapping. Teichmüller conjectured that the same would be true in all problems. For topological reasons, however, the conformal mappings cannot be single-valued mappings, except in the simplest cases, and by heuristic reasoning (we could almost say by slight of hand), he concluded that they must be determined, locally, by quadratic differentials with certain special properties. For two closed Riemann surfaces  $S_1$  and  $S_2$  the conjecture would be as follows: there exists a unique pair of analytic quadratic differentials  $\vartheta_1(z_1) dz_1^2$  and  $\vartheta_2(z_2) dz_2^2$ , and a positive constant  $k < 1$ , such that the extremal mapping can be expressed in local coordinates by

$$\sqrt{\vartheta_2} dz_2 = \sqrt{\vartheta_1} dz_1 + k \sqrt{\vartheta_1} d\bar{z}_1$$

After hard work Teichmüller finally managed to prove his conjecture, but it took several more years to produce a reasonably neat proof. Even now, the available proofs leave much to be desired in way of directness and simplicity.

The most remarkable feature of this result is the direct connection it establishes between quasiconformal mappings and quadratic differentials. This connection is entirely unexpected and is a sure sign that the role of quasiconformal mappings in function theory is far from superficial. Furthermore, any classically trained analyst cannot dismiss as pure coincidence that the number of linearly independent quadratic differentials is precisely equal to the number of parameters, or moduli, which according to Riemann characterize a closed Riemann surface of given genus. Teichmüller does not fail to make the most of this observation.

We now mention a slight generalization of Teichmüller's problem. If in the problem of polygons the vertices become everywhere dense, we are faced with the problem of a homeomorphic map of a disk on another with a given boundary correspondence. It is known under what conditions there exists a quasiconformal mapping with the given boundary correspondence, and Teichmüller's extremal problem makes sense. For a long time it was thought that this problem must have a unique solution, and that failure to prove it was only a sign of lacking technique. Quite recently, however, K. Strebel made a significant contribution to this problem, and, among other things, produced a beautifully simple counterexample. To describe this example, we remark that the disk can obviously be

*Figure 2*

replaced by any simply connected region. We choose to consider a region consisting of the lower half plane and a vertical half strip (Figure 2). Map this region onto itself by a stretching in the vertical direction. This produces a boundary correspondence which happens to be the identity on the horizontal part of the boundary. It is not difficult to show that the stretching is extremal for this correspondence. On the other hand we can find any number of mappings with the same boundary correspondence and the same maximal dilatation, for instance by use of the identity mapping in the lower half plane and the stretching in the "chimney." To obtain uniqueness it would thus be necessary to introduce some additional condition, and nobody knows what the proper condition is.

#### 4. THE GENERALIZED RIEMANN MAPPING THEOREM

We return to our earlier remark that a solution of the Beltrami equation  $f_{\bar{z}} = \mu f_z$  is a conformal mapping with respect to a Riemannian metric. If  $\mu$  is very regular, for instance analytic in the real sense, and defined in the whole plane, it is classical that the metric defines a structure of Riemann surface, and the general uniformization theorem asserts the existence of a conformal mapping on the whole plane or on the unit disk. As it happens, the quasiconformality rules out the latter case from the start.

The result is true without any regularity conditions whatsoever, except of course for the condition  $|\mu| \leq k < 1$  of quasiconformality, and it is this theorem that goes under the name of the generalized Riemann mapping theorem. The proof is implicit in the work of Morrey in 1938, but today we have a simpler proof based on the

theory of singular integrals of Calderon and Zygmund. In this form, the proof is due to Boyarskii and Vekua.

We sketch the proof under the simplifying assumption that  $\mu$  has compact support (this is a harmless restriction that is easy to remove). The mapping function is then analytic at  $\infty$ , and it may be normalized so that  $f(z) - z \rightarrow 0$  for  $z \rightarrow \infty$ . The familiar representation

$$(4.1) \quad f(\zeta) = -\frac{1}{\pi} \int \frac{f_z dx dy}{z - \zeta} + \zeta$$

is valid as soon as  $f_z \in L^p$  for some  $p > 2$ . Differentiation gives

$$f_{\zeta} = -\frac{1}{\pi} \int \frac{f_z dx dy}{(z - \zeta)^2} + 1$$

where the integral has to be interpreted as a Cauchy principal value. According to Calderon and Zygmund the integral transform may be regarded as a bounded operator in  $L^p$  with a norm  $C_p$  which tends to 1 for  $p \rightarrow 2$ . The Beltrami equation takes the form

$$\mu f_z = \mu T(\mu f_z) + \mu$$

The norm of  $\mu T$  is  $\leq kC_p$ , and therefore  $< 1$  for a suitable  $p > 2$ . We find that the Neumann series

$$f_z = \mu f_z = \mu + \mu T_\mu + \mu T_\mu T_\mu + \dots$$

converges in  $L^p$ , and substitution in (4.1) yields the desired solution.

It is perhaps interesting to note that we have not used the uniformization theorem. In fact, it is even possible to prove the uniformization by this method. With relatively simple proofs of the Calderon-Zygmund inequality available, this is a reasonable approach.

The result can be sharpened. Namely, if  $\mu$  depends differentiably or analytically on a parameter, then the solution  $f$ , normalized so as to be unique, will also depend differentially or analytically on the parameter.

Finally, instead of mapping the whole plane on itself we can map the unit disk on itself with a given complex dilatation  $\mu$ . The easiest way to arrive at this result is to extend  $\mu$  to the whole plane by means of an obvious reflexion process.

## 5. THE PROBLEM OF MODULI

The most important application of quasiconformal mappings is to the problem of moduli, for instance for compact Riemann surfaces of fixed genus  $g$ . Let us first describe Teichmüller's own approach, in particular for the most elementary case of  $g = 1$ .

All toruses are to be considered as points in a space, and the distance should depend on the maximal dilatation of the extremal quasiconformal mapping of one torus on another. The triangle inequality is satisfied if the distance is taken to be the logarithm of the least maximal dilatation. Consider the familiar representation of the torus as a period parallelogram. First, the shape of the torus depends only on the complex ratio  $\tau = \omega_2/\omega_1$  of the periods. Second, it depends on the period net rather than on the particular choice of a period parallelogram. In other words, the torus does not change when  $\tau$  is subjected to a modular transformation—a complication which it is preferable to avoid. Accordingly, we shall consider two period parallelograms as the same torus only if they are similar. With this convention the space of toruses is represented by the upper half plane, and we find that the Teichmüller distance is nothing but the noneuclidean distance between the corresponding points, a thoroughly satisfactory result.

For  $g > 1$  Teichmüller proposes a similar procedure. With deliberate oversimplification here, Teichmüller's approach is as follows. A point is not only a compact Riemann surface of genus  $g$ , but it is a surface together with a choice of generators of the fundamental group. A topological mapping automatically induces a correspondence of the generators, and the smallest maximum dilatation is to be determined in the class of mappings that induce the same correspondence. This defines the Teichmüller distance, and the metric space obtained in this manner is the Teichmüller space  $T_g$ . He proves that it is topologically equivalent to  $R^{6g-6}$ . As a topological theorem this is a rediscovery of a result of Fricke, or so it would seem, for Fricke is almost undecipherable for modern mathematicians.

## 6. THE COMPLEX STRUCTURE

It is not for lack of admiration that this author has given such a brief account of Teichmüller's theory of moduli. The reason is

that his method has been superseded by simpler and more efficient procedures. Curiously enough, it was by abandoning the idea of extremal quasiconformal mappings that further progress became possible. Interesting and important as it is, the extremal problem seems to lead very rapidly to serious technical difficulties, and the overcoming of these difficulties is not essential for the problem of moduli.

The metrical structure obtained by Teichmüller is closer to a Finsler than to a Riemann metric. Actually, the classical problem calls for a complex structure, and Teichmüller was well aware that his method did not yet solve this problem. In fact, he refutes his own method and says explicitly that it is no good for the purpose of establishing a complex structure. This prediction turned out to be wrong.

In a final effort Teichmüller produced a solution of the structure problem, by an entirely different method, but it was so cumbersome that it is doubtful whether anybody else has checked all the details. Many years later E. Rauch and this author found, independently of each other that the original method, used primarily as a convenient parametrization, does lead fairly easily to the existence of a natural complex structure on the space  $T_g$ . The most significant development, however, is due to L. Bers. He found, as has been indicated, that Teichmüller's extremal problem can be dispensed with altogether, and that the most direct approach is via a closer study of the Beltrami equation. In other words, it is the full theory of quasiconformal mappings, rather than a special problem, which is the right clue to the moduli of Riemann surfaces.

It is only fair to mention, at this point, that the algebraists have also solved the problem of moduli, in some sense even more completely than the analysts. Because of the different language, it is at present difficult to compare the algebraic and analytic methods, but it would seem that both have their own advantages. We shall treat only the analytic problem.

Let  $W_0$  be a closed Riemann surface of genus  $g > 0$ . The universal covering surface of  $W_0$  can be mapped conformally on the upper half plane  $H$  with the cover transformations corresponding to a discontinuous group  $\Gamma_0$  of Möbius transformations, a Fuchsian group. Clearly,  $W_0$  can be identified with the quotient space  $H/\Gamma_0$ . If  $W$  is another surface of the same genus, the corresponding group  $\Gamma$  is isomorphic to  $\Gamma_0$ , and two such  $W$ 's are conformally

equivalent if and only if the groups  $\Gamma$  are conjugate subgroups of the full group of noneuclidean motions. The conformally different Riemann surfaces are thus identified with conjugacy classes. A point in Teichmüller space, however, will be represented by a group  $\Gamma$  together with an isomorphism  $\Theta$  of  $\Gamma^0$  on  $\Gamma$ , again with the understanding that two isomorphisms are regarded as equivalent if they differ by an inner automorphism. More precisely, the space we obtain is the Teichmüller space  $T(W_0)$  with  $W_0$  as a reference point, but it is perfectly obvious how to change the reference point.

To study the structure of  $T(W_0)$  we need an analytic representation. For this purpose we consider the linear space  $B(\Gamma^0)$  of Beltrami differentials: its elements are measurable bounded functions  $\mu$ , defined on  $H$ , such that the expression  $\mu d\bar{z}/dz$  is invariant under  $\Gamma^0$ . Assume, in particular, that  $\|\mu\|_\infty < 1$ . By the generalized Riemann mapping theorem there exists a solution of the Beltrami equation  $f_z = \mu f_{\bar{z}}$  which maps  $H$  homeomorphically onto itself. If  $A \in \Gamma^0$  it follows at once that  $f(Az)$  is another solution, and hence that  $f(Az) = A''[f(z)]$  where  $A''$  is conformal, and consequently a Möbius transformation. The mapping  $\Theta'' : A \rightarrow A''$  is an isomorphism which determines a group  $\Gamma''$  and a point of  $T(W_0)$ . Because any two closed Riemann surfaces of the same genus are quasiconformally equivalent, it is easy to see that all points of  $T(W_0)$  can be obtained in this way. Different  $\mu$  may determine the same point, however, and it becomes essential to investigate when this is so.

Let the mapping function associated with  $\mu$  be normalized so that 0, 1,  $\infty$  remain fixed, and let it be denoted by  $f^\mu$ . It is easily seen that  $\mu_1$  and  $\mu_2$  determine the same Teichmüller point if and only if  $f^{\mu_1} = f^{\mu_2}$  on the real axis. From a global point of view this condition is rather unmanageable, but the corresponding infinitesimal condition is quite simple. Because the reference point  $W_0$  is arbitrary, it is sufficient to study the behavior of  $f^\mu$  for small  $\mu$ . For any  $\nu \in B(\Gamma^0)$  we introduce the Fréchet derivative

$$f'[\nu] = \lim_{t \rightarrow 0} \frac{f^{t\nu} - z}{t}$$

which is known to exist. Clearly,  $\nu$  produces an infinitesimally vanishing change if and only if  $f'[\nu] = 0$  on the real axis. This is a linear condition on  $\nu$ , and a simple computation which we shall not

reproduce shows that it can be expressed by

$$(6.1) \quad \int_{\Omega/\Gamma^0} \nu \varphi \, dx \, dy = 0$$

where  $\varphi$  runs through all quadratic differentials, that is, all analytic functions such that  $\varphi \, dz^2$  is invariant under  $\Gamma^0$ . Let  $N(\Gamma^0)$  denote the subspace of  $B(\Gamma^0)$  for which (6.1) is satisfied. Because there are  $3g - 3$  linearly independent quadratic differentials, the quotient space  $B(\Gamma^0)/N(\Gamma^0)$  has  $3g - 3$  complex dimensions. If  $\nu_1, \dots, \nu_{3g-3}$  belong to linearly independent cosets, a simple application of the implicit function theorem shows that the Teichmüller points near the origin can be uniquely represented by  $\mu = \tau_1 \nu_1 + \dots + \tau_{3g-3} \nu_{3g-3}$  where  $\tau_1, \dots, \tau_{3g-3}$  are complex parameters. A somewhat deeper study shows that these parameters define a complex structure on  $T_g$ . This neat derivation of the complex structure is due to L. Bers.

To complete the picture, let us mention that there is a natural Riemannian metric on  $T_g$ . In fact, the tangents are in one-to-one correspondence with the quadratic differentials, and it suffices to introduce a Hermitian inner product on the space of quadratic differentials. Such a product was defined, years ago, by H. Petersson, in the theory of automorphic forms. Surprisingly enough, quite a few properties of the Riemannian metric can be proved by explicit calculations. For instance, it is Kählerian, and the analytic sectional curvatures are negative. An important open problem is the connection between the Riemannian structure and the Teichmüller metric.

## 7. THE GLOBAL APPROACH

There is a more direct method which permits us to introduce the complex structure in a single step, without passing to infinitesimal derivations.

Given a Beltrami differential  $\mu$  we constructed  $f^\mu$  as a mapping of the upper half plane on itself with complex dilatation  $\mu$ . To obtain the mapping,  $\mu$  is extended to the lower half plane by reflection, and this defines  $f^\mu$  by the generalized Riemann mapping theorem as a mapping of the whole plane upon itself. Nothing prevents us, however, from using another extension of  $\mu$ , for instance to set  $\mu = 0$  in the lower half plane. The corresponding mapping func-

tion  $f_\mu$  is analytic in the lower half plane; thus we obtain a quasiconformal mapping of  $U$  on a Jordan region  $\Omega$  together with a conformal mapping of the lower half plane  $U^*$  on its complement  $\Omega^*$ . Moreover, if we set  $f_\mu(Az) = A_\mu[f_\mu(z)]$ , then  $A_\mu$  is still a linear transformation, and we obtain a new group  $\Gamma_\mu$ , isomorphic to  $\Gamma^0$ , which is discontinuous in the invariant regions  $\Omega$  and  $\Omega^*$ . Such groups are called quasi-Fuchsian.

Because  $f_\mu$  is determined only up to a linear transformation, it is natural to pass to its Schwarzian derivative, usually denoted by  $\{f_\mu, z\}$ , defined in the lower half plane. The interesting observation is that this turns out to be a quadratic differential  $\varphi_\mu$  with respect to  $\Gamma^0$ . Moreover,  $\varphi_\mu$  completely determines the values of  $f_\mu$  on the real axis, up to a linear transformation, and thereby the group  $\Gamma_\mu$ . It follows that  $T(W_0)$  can be represented as a subset of  $Q(\Gamma^0)$ , the space of quadratic differentials, and hence as a set in  $C^{3g-3}$ . This set is open and bounded. The boundedness is an easy consequence of a theorem of Z. Nehari, but the openness is not so elementary, and at present it cannot be proved directly except by use of infinitesimal deformations. The proof shows that the complex structure induced by the imbedding is the same that we have already introduced.

As a bounded region  $T_g$  has a Bergman kernel function and a Riemannian metric deduced from the kernel function. It is not known how the Bergman metric is related to the Petersson metric. The nature of the boundary and its role for the problem of moduli is also an open question.

Although we have not been very explicit, it should be clear from what has been said that  $T_g$  admits a group  $M_g$  of isometries with the property that Teichmüller points which correspond to each other under such an isometry represent conformally equivalent Riemann surfaces. We call  $M_g$  the modular group, because it is the exact analogue of the modular group in the half plane. The quotient space  $T_g/M_g$  is the space of Riemann surfaces of genus  $g$ . It is a complex space, but not a manifold, for it has singularities at the fixpoints of elements of the group  $M_g$ . The fixpoints occur at points which correspond to surfaces with conformal self-mappings, and E. Rauch has determined all cases where the singularities are nonremovable.

# 6

## Differential Topology

*J. Milnor*

The basic objects studied in differential topology are smooth manifolds, sometimes with boundary, and smooth mappings between such manifolds. Here the word "smooth" is used to mean "differentiable of class  $C^\infty$ ." To give a rough idea of the flavor of this field, let us list a few of its central problems.

*The Diffeomorphism Problem.* Given two smooth manifolds  $M$  and  $M'$ , how can we decide whether or not there is a diffeomorphism from  $M$  to  $M'$  (i.e., a smooth homeomorphism with smooth inverse)?

*The Cobordism Problem.* Given a smooth compact manifold  $M$  without boundary, does there exist a smooth compact manifold  $W$  whose boundary is equal to  $M$ ? We may refine this problem by putting extra structures on both  $M$  and  $W$ . For example, we can require an orientation, or a weakly complex structure, or we may require that  $M$  and  $W$  should be  $k$ -connected (compare [10]).

*The Imbedding Problem.* Given  $M$  and  $M'$ , does there exist a smooth imbedding  $M \rightarrow M'$ ? If so, can we classify all such imbeddings? For example, the problem of classifying imbeddings of the circle in 3-space forms the field of "knot theory."

CHARACTERIZATIONS OF THE  $n$ -SPHERE

Let us single out one particular case of the diffeomorphism problem for consideration, namely the problem of characterizing the  $n$ -sphere. The many different tools which can be brought to bear on this one question will provide a survey of much of the field of differential topology.

First let us ask what conditions on a smooth manifold guarantee that it is *homeomorphic* to  $S^n$ .

The first such characterization is due to Reeb [16]. Let  $M$  be a smooth  $n$ -dimensional manifold. By a *Morse function*  $f$  on  $M$  will be meant a smooth real valued function whose critical points are all nondegenerate. Thus in a neighborhood of each critical point we can choose local coordinates  $u_1, \dots, u_n$  so that  $f = \text{constant} \pm u_1^2 \pm \dots \pm u_n^2$ . (Morse [12, Lemma 4] or [11, §2.2].)

*Theorem 1.* *If  $M$  is compact, without boundary, and possesses a Morse function with only two critical points, then  $M$  is homeomorphic to  $S^n$ .*

OUTLINE OF PROOF. (See Figure 1.) Let  $m_0, m_1$  denote the minimum and maximum of  $f$ . Thus in a neighborhood of its minimum point we have

$$f = m_0 + u_1^2 + \dots + u_n^2$$

It follows that the set of points  $x$  where  $f(x) \leq m_0 + \epsilon$  is diffeomorphic to the disk  $D^n$ . Similarly, the set of  $x$  with  $f(x) \geq m_1 - \epsilon$

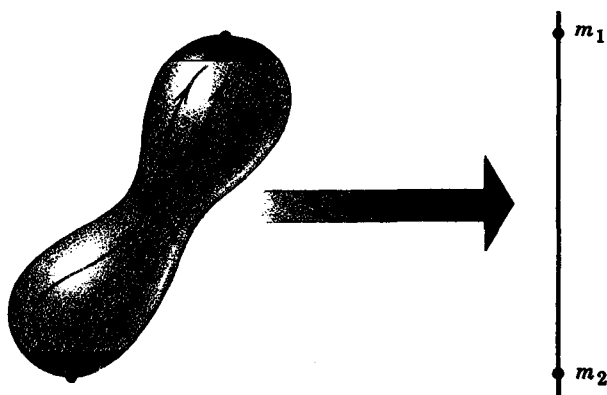


Figure 1

is diffeomorphic to  $D^n$ . But, deforming  $M$  into itself along the gradient lines of the function  $f$  [i.e., along the orthogonal trajectories of the surfaces  $f^{-1}$  (constant)] we can slide the disk,  $f \leq m_0 + \epsilon$ , up so that it covers precisely the disk  $f \leq m_1 - \epsilon$ . (Compare [7] or [11, §§3, 4].) Thus  $M$  can be considered as the union of two imbedded disks which intersect only along their common boundary. But this implies that  $M$  is homeomorphic to  $S^n$ . To see this, consider first the following.

*Lemma 1. Any homeomorphism  $h: S^{n-1} \rightarrow S^{n-1}$  extends to a homeomorphism  $H: D^n \rightarrow D^n$ .*

PROOF. Set  $H(tu) = th(u)$  for  $0 \leq t \leq 1$ , where  $u$  denotes an arbitrary unit vector.

Now if  $M$  is the union of two topologically imbedded disks  $g_0(D^n)$ ,  $g_1(D^n)$  which intersect precisely along their common boundary, we can first choose any homeomorphism from  $g_0(S^{n-1}) = g_1(S^{n-1})$  to the equator of  $S^n$ , and then extend, using Lemma 1, to a homeomorphism which maps  $g_0(D^n)$  to the southern hemisphere and  $g_1(S^{n-1})$  to the northern hemisphere. This completes the proof.

*Remark.* Note that the differentiable structure is destroyed in the course of this proof. The reason is that there is no differentiable analogue of Lemma 1. (Even if  $h: S^{n-1} \rightarrow S^{n-1}$  is a diffeomorphism, the extension we have constructed is highly nondifferentiable at the origin, unless  $h$  happens to be orthogonal.) We will return to this point later, in Section 3.

Following is another partial characterization of  $S^n$  (compare [17, 8]).

*Theorem 2. Let  $M$  be a compact smooth manifold which is the union of two open sets, each diffeomorphic to a euclidean space. Then  $M$  is homeomorphic to a sphere.*

In fact the proof will show that  $M$  with a single point deleted is diffeomorphic to the euclidean space  $R^n$ . The proof is based on two lemmas, both of which are interesting in their own right.

*Lemma 2 (Palais and Cerf). Let  $f_1$  and  $f_2$  be smooth, orientation preserving imbeddings of the disk  $D^n$  into the interior of a connected manifold  $M^n$ . Then there exists a diffeomorphism  $h$  of  $M^n$  onto itself so that  $h \circ f_1 = f_2$ .*

For the proof, which is quite elementary, the reader is referred to [3] or [14].

**Lemma 3 (Brown, Stallings).** *Let  $M$  be a paracompact manifold such that every compact subset is contained in an open set diffeomorphic to euclidean space. Then  $M$  itself is diffeomorphic to euclidean space.*

**PROOF.** (Compare [2, 21].) It is not difficult to show that  $M$  is a monotone union of disks. That is, we can find submanifolds with boundary

$$W_1 \subset W_2 \subset W_3 \cdots \subset M$$

with union  $M$  so that each  $W_i$  is diffeomorphic to  $D^n$ , and so that each  $W_i$  is contained in the interior of  $W_{i+1}$ . We wish to compare this sequence with the sequence

$$D_1^n \subset D_2^n \subset D_3^n \subset \cdots \subset R^n$$

where  $D_i^n$  denotes the disk of radius  $i$  in euclidean space. Start with any diffeomorphism  $f_1: D_1^n \rightarrow W_1$ . Using Lemma 2, this can be extended to a diffeomorphism  $f_2: D_2^n \rightarrow W_2$ , and so on. (To see this, consider the following diagram

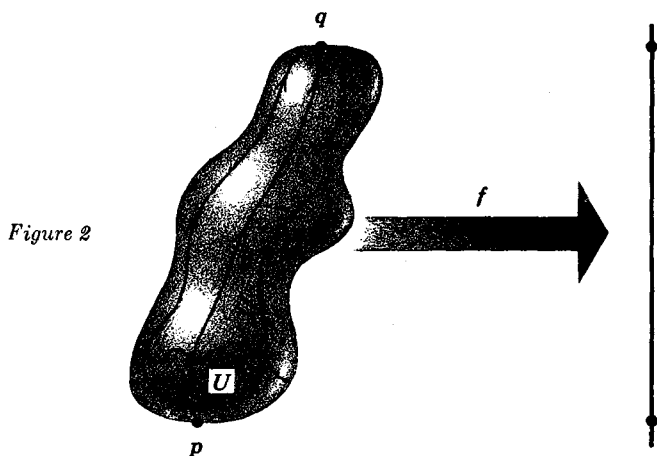
$$\begin{array}{ccc} W_i \subset W_{i+1} & & \\ \uparrow f_i & \nwarrow h & \\ D_i^n \subset D_{i+1}^n & \nearrow g & W_{i+1} \end{array}$$

where  $g$  is an arbitrary orientation preserving diffeomorphism. Choosing a diffeomorphism  $h$  as in Lemma 2, we can now set  $f_{i+1}$  equal to  $h \circ g$ .) Finally, piecing together all these diffeomorphisms  $f_i$ , we obtain the required diffeomorphism  $R^n \rightarrow M$ .

Using these lemmas we can prove a sharpened form of Theorem 1. Let  $M$  be compact, without boundary, and let  $f: M \rightarrow R$  be a smooth function with only two critical points.

**Theorem 1'.** *Even if these critical points are allowed to be degenerate, it still follows that  $M$  is homeomorphic to  $S^n$ .*

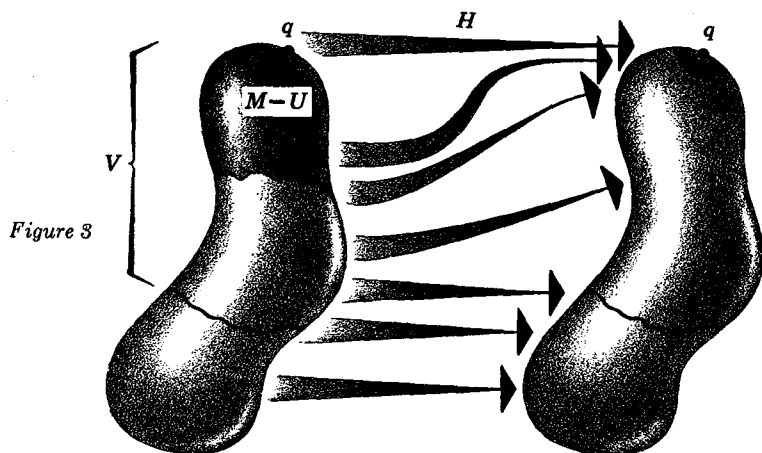
**PROOF.** (See Figure 2.) Let  $p$  and  $q$  be the critical points and let  $U$  be a neighborhood of  $p$  which is diffeomorphic to  $R^n$ , with  $q \notin U$ . By deforming  $M$  into itself along the gradient lines of  $f$ , we can stretch  $U$  so that it covers any compact subset of  $M - q$ .



Hence it follows from Lemma 3 that  $M - q$  is diffeomorphic to  $R^n$ , which clearly completes the proof.

The proof of Theorem 2 is similar. Let  $M$  be covered by open sets  $U$  and  $V$  which are diffeomorphic to  $R^n$  (Figure 3). Given  $q \in M - U \subset V$  we will show that any compact subset of  $M - q$  can be covered by an open set diffeomorphic to  $U$ .

Since  $V$  is diffeomorphic to  $R^n$ , it is easy to construct a diffeomorphism  $h: V \rightarrow V$  which (1) satisfies  $h(q) = q$ , (2) shrinks the



compact set  $M - U$  down into an arbitrarily small neighborhood of  $q$ , and (3) coincides with the identity outside of a larger compact set. It follows from (3) that  $h$  extends to a diffeomorphism  $H$  of  $M$ . Clearly any compact subset of  $M - q$  can be covered by  $H(U)$  for suitable choice of the diffeomorphism  $H$ . Therefore it follows from Lemma 3 that  $M - q$  is diffeomorphic to  $R^n$ . This completes the proof of Theorem 2.

One of the most striking properties of the sphere  $S^n$  is that the complement of each point is contractible.

*Problem 1. Given a smooth manifold  $M$  such that  $M - p$  is contractible, does it follow that  $M$  is homeomorphic to a sphere?*

It follows without too much difficulty from this hypothesis that  $M$  is compact and has the homotopy type of a sphere. Conversely, if  $M$  is compact, without boundary, and has the homotopy type of a sphere, it can be shown that  $M - p$  is contractible.

If the dimension  $n$  of  $M$  is 0, 1, or 2, then  $M$  must actually be diffeomorphic to  $S^n$ . Here is a proof for  $n = 2$ . Choose a Riemannian metric and an orientation for  $M$ . By a classical theorem (the existence of "isothermal coordinates") each small neighborhood in  $M$  can be mapped conformally and diffeomorphically onto a region in the plane. Thus  $M$  becomes a Riemann surface. Since  $M$  is simply connected, the classical "uniformization theorem" asserts that  $M$  is conformally diffeomorphic to either the complex plane, the open unit disk, or the Gauss 2-sphere. But only the last possibility satisfies our hypothesis.

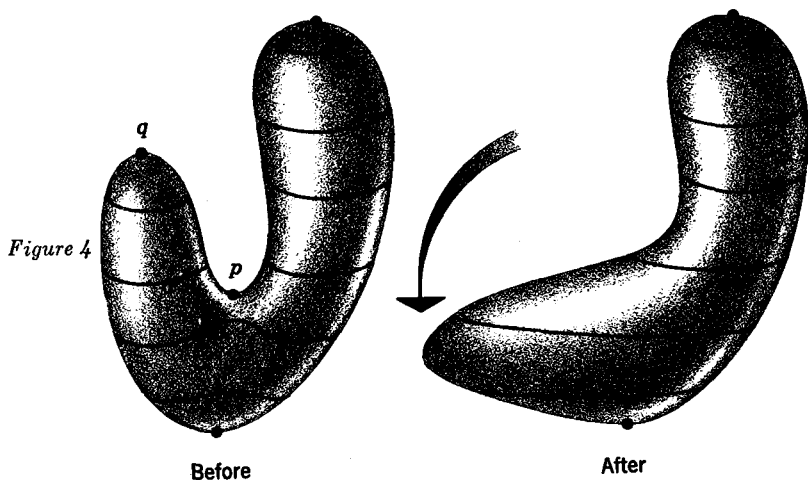
If the dimension  $n$  is 3, we have the classical Poincaré problem. So far, all attempts to solve this problem have foundered. For  $n = 4$  the problem is also unsolved.

For  $n \geq 5$  the problem has been solved affirmatively by Stallings and Zeeman [20, 24] and by Smale. In particular we have the following.

*Theorem of Smale [19]. If  $M$  is a smooth homotopy  $n$ -sphere of dimension  $n \geq 5$ , then  $M$  admits a Morse function with only two critical points.*

Hence Theorem 1 implies that  $M$  is homeomorphic to the  $n$ -sphere.

The proof of this theorem is much more difficult than anything we have encountered so far. The basic idea can be outlined very



roughly as follows. It is not difficult to construct a Morse function  $f: M \rightarrow \mathbb{R}$  with many critical points. Furthermore  $f$  can be chosen so that the critical points occur in the proper order, in the following sense. If

$$f(u_1, \dots, u_n) = \text{constant} - u_1^2 - \dots - u_\lambda^2 + u_{\lambda+1}^2 + \dots + u_n^2$$

in terms of local coordinates  $u_1, \dots, u_n$  near a critical point  $p$ , then the integer  $\lambda = \lambda(p)$  is called the *index* of this critical point. Now  $f$  can be chosen so that  $f(p) = \lambda(p)$  for each critical point  $p$ . Thus the minimum points, with  $\lambda = 0$ , occur in  $f^{-1}(0)$ , the maximum points occur in  $f^{-1}(n)$ , and the remaining critical points come in between.

The difficult part of the proof now consists in showing that a critical point of index  $\lambda$  and a critical point of index  $\lambda + 1$  can sometimes be mutually cancelled. Thus in Figure 4 the critical point  $p$  of index 1 and the critical point  $q$  of index 2 can be mutually eliminated, by suitably changing the function. Repeating this procedure over and over, we eventually eliminate all critical points which are not essential in order to give the manifold  $M$  its proper homology groups. But a homotopy sphere has homology only in dimensions 0 and  $n$ . Hence we are left with only two critical points, with indices 0 and  $n$  respectively.

## 2. SOME EXOTIC SPHERES

This section will construct an example of a manifold which is homeomorphic, but not diffeomorphic, to a sphere. The proof will be based on the Hirzebruch signature theorem.

Let  $V$  be a compact oriented manifold of dimension  $4k$ . The signature (or index\*)  $\sigma(V)$  is defined as follows, following Hermann Weyl [22, p. 41]. Let  $\alpha, \beta \in H_{2k}(V; \mathbb{Z})$  be homology classes with integer coefficients. Then the intersection number  $\alpha \cdot \beta$  is a well-defined integer. This intersection number is symmetric in  $\alpha$  and  $\beta$  since the dimension  $2k$  is even. Passing to real coefficients, we can choose a basis  $\alpha_1, \dots, \alpha_r$  for the vector space  $H_{2k}(V; \mathbb{R})$  so that the matrix of intersection numbers  $(\alpha_i \cdot \alpha_j)$  is diagonal. Now the number of positive diagonal entries minus the number of negative diagonal entries is called the *signature*  $\sigma$  of  $V$ .

The following fundamental observation is due to Thom. *Suppose that  $V$  is the boundary of a compact oriented manifold  $W^{4k+1}$ . Then the signature  $\sigma(V)$  is zero.* Thom's proof is based on the Poincaré duality theorem.

We will also need to make use of the *Pontrjagin classes* of a smooth manifold  $M$ . Without attempting a definition, here are some basic properties.

1. To each smooth manifold  $M$  there are associated cohomology classes  $p_i \in H^{4i}(M; \mathbb{Z})$  for  $i = 1, 2, 3, \dots$
2. If  $U$  is an open subset of  $M$ , then  $p_i(U)$  is equal to  $p_i(M)$  restricted to  $U$ .
3. If  $M$  is parallelizable, then  $p_i(M) = 0$ .

Now let  $V$  be a closed oriented manifold of dimension  $4k$ . If  $i_1 + i_2 + \dots + i_r = k$ , then the cohomology class  $p_{i_1} p_{i_2} \dots p_{i_r} \in H^{4k}(V; \mathbb{Z})$  gives rise to an integer which is denoted by  $p_{i_1} \dots p_{i_r}[V]$ . These integers are called the *Pontrjagin numbers* of  $V$ .

If  $V$  is a boundary then these Pontrjagin numbers are all zero. Conversely, we have the following:

*Thom Cobordism Theorem.* *If the Pontrjagin numbers  $p_{i_1} p_{i_2} \dots p_{i_r}[V]$  are all zero, then the  $m$ -fold disjoint sum  $\dot{V} + V + \dots + V$  is the boundary of a compact oriented manifold, for some  $m > 0$ .*

\* The term "index" is preferred by Hirzebruch and others. I have substituted "signature" to avoid confusion with the Morse index, as used in Section 1.

An important consequence is the following.

*Hirzebruch Signature Theorem.* The signature of any closed, oriented  $4k$ -manifold can be expressed as a linear combination of its Pontrjagin numbers, where the coefficients are rational numbers which depend only on the dimension. In particular:

$$\sigma(V^4) = \frac{p_1[V^4]}{3}$$

$$\sigma(V^8) = \frac{7p_2[V^8] - p_1^2[V^8]}{45}$$

$$\sigma(V^{12}) = \frac{62p_3[V^{12}]}{945} + \dots$$

$$\sigma(V^{16}) = \frac{127p_4[V^{16}]}{4725} + \dots$$

$$\sigma(V^{20}) = \frac{146p_5[V^{20}]}{13365} + \dots$$

where the dots indicate terms in  $p_1, \dots, p_{k-1}$ .

PROOF. Let  $V$  and  $V'$  be two manifolds with the same Pontrjagin numbers. Form  $V - V'$  (the disjoint sum in which the orientation of  $V'$  has been reversed). Then all Pontrjagin numbers of  $V - V'$  are zero. Hence some multiple

$$(V - V') + \dots + (V - V')$$

is a boundary. This implies that the signature

$$\sigma((V - V') + \dots + (V - V')) = m\sigma(V) - m\sigma(V')$$

is zero, and hence that  $\sigma(V) = \sigma(V')$ .

Thus  $\sigma(V)$  is a function of the Pontrjagin numbers of  $V$ . Since both signature and Pontrjagin numbers are integers which behave additively when we form disjoint sums, it is clear that this function must be linear with rational coefficients.

The explicit formulas for  $k = 1, 2, 3, 4, 5$  can be obtained by computing the signature and Pontrjagin numbers for suitable examples, and then solving the resulting linear equations. For further details the reader is referred to Hirzebruch [4].

For our purposes we will only need the following. A manifold

will be called *almost parallelizable* if it becomes parallelizable when a single point is removed.

*Corollary.* The signature of a closed, almost parallelizable manifold of dimension 8 is always divisible by 7. Similarly, for a closed, almost parallelizable manifold of dimension 12, 16 or 20, the signature is divisible by 62, 127, or 146 respectively.

**PROOF.** We have  $p_i(V^{4k} - x) = 0$ , hence  $p_i(V^{4k}) = 0$  for  $i < k$ . Thus the signature theorem reduces to

$$\sigma(V^8) = \frac{7p_2[V^8]}{45} \equiv 0 \pmod{7}$$

$$\sigma(V^{12}) = \frac{62p_3[V^{12}]}{945} \equiv 0 \pmod{62}$$

and so on.

In higher dimensions we obtain analogous results. (However, sharper results in this direction can be obtained by a different method: see [5].)

*Lemma 4.* For  $k \geq 2$  there exists a compact parallelizable  $4k$ -dimensional manifold  $W$  with signature  $+8$ , such that the boundary of  $W$  is a homotopy sphere.

**PROOF.** We will construct  $W$  in such a way that the matrix of intersection numbers is as follows:

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 \end{bmatrix}$$

This remarkable matrix was suggested to the author by Hirzebruch. Note that it is positive definite, with determinant  $+1$ , and has only even entries on the diagonal.

As basic building block for the manifold  $W$  we take a tubular neighborhood  $T$  of the diagonal in  $S^{2k} \times S^{2k}$ . The homology group  $H_{2k}(T; \mathbb{Z})$  is infinite cyclic, and the intersection number of a gener-

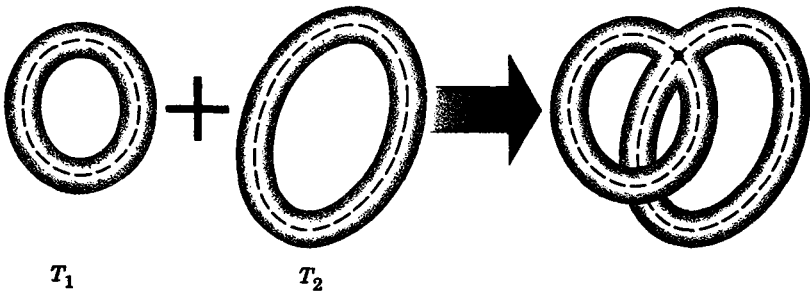


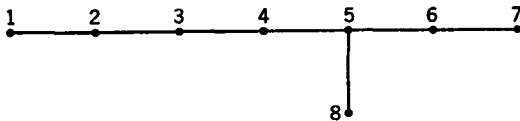
Figure 5

ator with itself is  $+2$ . [PROOF: if  $\alpha, \beta \in H_{2k}(S^{2k} \times S^{2k})$  are the obvious generators, then  $(\alpha + \beta) \cdot (\alpha + \beta) = 2\alpha \cdot \beta = 2$ .]

Proof that  $T$  is parallelizable. Note that  $S^{2k} \times S^{2k}$  can be imbedded in  $R^{4k+1}$  (as the boundary of a neighborhood of  $S^{2k}$ ). Hence its tangent bundle is induced from that of the unit sphere by means of the Gauss map  $g: S^{2k} \times S^{2k} \rightarrow S^{4k}$ . Since  $g|T$  is homotopic to a constant, it follows that  $T$  is parallelizable.

Next we introduce the operation of "plumbing together" two copies  $T_1$  and  $T_2$  of  $T$ . By this we mean the operation of matching together a region in  $T_1$  and a region in  $T_2$  in such a way that the central  $2k$ -sphere of  $T_1$  will have intersection number  $+1$  with the central  $2k$ -sphere of  $T_2$ . We must then "round off" the corners, so as to obtain a smooth manifold with boundary. For the 2-dimensional case this operation is illustrated in Figure 5.

In practice we need not two but eight copies of  $T$ . These are to be plumbed together according to the following schema:



That is,  $T_2$  is to be plumbed together with  $T_1$  and  $T_3$ ;  $T_5$  is to be plumbed together with  $T_4$ ,  $T_6$ , and  $T_8$ ; and so on. In this way we obtain a smooth manifold  $W$  which clearly has the following properties:

1. The manifold  $W$  has the homotopy type of a bouquet consisting of eight copies of  $S^{2k}$  which intersect at a single point.

2. If  $\alpha_1, \dots, \alpha_8 \in H_{2k}(W; \mathbb{Z})$  is the corresponding homology basis, then the matrix  $(\alpha_i \cdot \alpha_j)$  of intersection numbers is as indicated previously, positive definite with determinant  $+1$ .

Thus it follows that the signature of  $W$  is  $+8$ .

3. The boundary  $\partial W$  is a homology sphere.

For the Poincaré duality theorem implies that

$$H_i(W, \partial W; \mathbb{Z}) = \begin{cases} \mathbb{Z} + \mathbb{Z} + \mathbb{Z} + \mathbb{Z} + \mathbb{Z} + \mathbb{Z} + \mathbb{Z} + \mathbb{Z} & \text{for } i = 2k \\ \mathbb{Z} & \text{for } i = 4k \\ 0 & \text{otherwise} \end{cases}$$

Furthermore the homomorphism

$$H_{2k}(W; \mathbb{Z}) \rightarrow H_{2k}(W, \partial W; \mathbb{Z})$$

corresponds to our intersection matrix, and hence is an isomorphism. Now the homology exact sequence of the pair  $(W, \partial W)$  shows that  $H_*(\partial W; \mathbb{Z}) \cong H_*(S^{4k-1}; \mathbb{Z})$ .

4.  $\partial W$  is simply connected.

PROOF. Any circle in  $\partial W$  bounds a disk in  $W$ . By a general position argument, since  $2k + 2 < 4k$ , this disk can be pushed off of the central  $2k$ -spheres of the tubes  $T_i$ , and hence can be pushed into the boundary of  $W$ .

It now follows, by standard arguments in homotopy theory, that  $\partial W$  is a homotopy sphere. Since  $W$  is clearly parallelizable, this completes the proof of Lemma 4.

*Theorem 3. The manifold  $\partial W$  is homeomorphic to the sphere  $S^{4k-1}$  but is not diffeomorphic to  $S^{4k-1}$ .*

PROOF. It follows immediately from Smale or Stallings that the homotopy sphere  $\partial W$  is homeomorphic to  $S^{4k-1}$ . (Recall that the dimension  $4k - 1$  is  $\geq 7$ .)

We will only prove the second statement for the special cases  $k = 2, 3, 4, 5$ . Suppose that  $\partial W$  were diffeomorphic to  $S^{4k-1}$ . Then by pasting a  $4k$ -disk onto the boundary of  $W$  we would obtain a smooth closed manifold  $V = W \cup D^{4k}$ . Clearly  $V$  is almost parallelizable. Thus, according to the corollary given earlier, the signature of  $V$  must be a multiple of 7 or of 62 or 127 or 146 respectively. But  $\sigma(V) = \sigma(W) = 8$ . This contradiction completes the

proof for  $2 \leq k \leq 5$ . For higher values of  $k$  the reader is referred to [6].

### 3. THE GROUP $\Gamma_n$

Since we have seen that exotic spheres exist, it is natural to try to classify them. For this purpose we introduce an abelian group  $\Gamma_n$  which can be described in two different ways.

*First Description.* Let  $\text{Diff}^+(D^n)$  be the group of all orientation preserving diffeomorphisms from a closed  $n$ -disk onto itself, and let  $\text{Diff}^+(S^{n-1})$  be the corresponding group of diffeomorphisms of its boundary. Consider the restriction homomorphism

$$r: \text{Diff}^+(D^n) \rightarrow \text{Diff}^+(S^{n-1})$$

We assert that the image of this homomorphism is a normal subgroup. Hence the quotient group

$$\Gamma_n = \frac{\text{Diff}^+(S^{n-1})}{r \text{Diff}^+(D^n)}$$

is defined.

*Second Description.* A closed oriented manifold  $M$  will be called a *twisted sphere* if it admits a Morse function with two critical points. Let  $\Gamma'_n$  denote the set of all oriented-diffeomorphism classes of twisted  $n$ -spheres. To make this into a group we introduce the *connected sum* operation.

Given connected oriented  $n$ -manifolds  $M_1, M_2$ , choose imbeddings

$$f_1: D^n \rightarrow M_1, \quad f_2: D^n \rightarrow M_2$$

To make the orientations come out right it is important that one of these two imbeddings should preserve orientation and the other one should reverse orientation. The connected sum  $M_1 \# M_2$  is now formed from

$$[M_1 - f_1(\text{Int } D^n)] \cup [M_2 - f_2(\text{Int } D^n)]$$

by pasting together the two boundaries under the diffeomorphism

$$f_2 \circ f_1^{-1}: f_1(S^{n-1}) \rightarrow f_2(S^{n-1})$$

Better still, in order to make the differentiable structure clear, extend  $f_1$  and  $f_2$  to imbeddings  $F_i: R^n \rightarrow M_i$ . Then  $M_1 \# M_2$  can

be formed from

$$[M_1 - F_1(0)] \cup [M_2 - F_2(0)]$$

by identifying each  $F_1(x)$  with  $F_2(x/\|x\|^2)$ . (Note that this correspondence preserves orientation.) Applying Lemma 2 to  $F_i$  restricted to some large disk, we can prove that  $M_1 \# M_2$  is well defined up to orientation preserving diffeomorphism. (Compare [8, 6].)

To compare the group  $\Gamma_n$  with  $\Gamma'_n$  we will construct a homomorphism

$$\text{Diff}^+(S^{n-1}) \rightarrow \Gamma'_n$$

Given  $h \in \text{Diff}^+(S^{n-1})$  let  $M(h)$  be the twisted sphere which is obtained from two copies of  $R^n$  by identifying each  $x \in R^n - 0$  in the first with  $y = h(x/\|x\|)/\|x\|$  in the second. Then

$$f = \frac{\|x\|^2}{1 + \|x\|^2} = \frac{1}{1 + \|y\|^2}$$

is a Morse function with only two critical points. Taking the orientation of  $M(h)$  from the first copy of  $R^n$  we obtain a well-defined twisted sphere.

It is easily verified that this construction defines a homomorphism from  $\text{Diff}^+(S^{n-1})$  onto  $\Gamma'_n$ . (Compare [7, p. 402].) Let us look at the kernel. Suppose that there is an orientation preserving diffeomorphism  $g$  from  $M(h_1)$  to  $M(h_2)$ . According to Lemma 2 we may assume that  $g$  carries the point of  $M(h_1)$  with coordinate  $x$  to the point of  $M(h_2)$  with the same coordinate  $x$  for all  $\|x\| \leq 1$ . In terms of the  $y$  coordinates, this means that  $g$  carries the point with coordinate  $y = \text{th}_1(u)$  to the point with coordinate  $y = \text{th}_2(u)$  for all  $t \geq 1$  and all  $u \in S^{n-1}$ . Thus we have a diffeomorphism  $R^n \rightarrow R^n$  which takes  $S^{n-1}$  into itself by the diffeomorphism  $h_2 \circ h_1^{-1}$ . Hence  $h_2 \circ h_1^{-1}$  belongs to the image

$$r \text{Diff}^+(D^n) \subset \text{Diff}^+(S^{n-1})$$

Conversely it can be shown that any element  $h|S^{n-1}$  of the image gives rise to a manifold  $M(h|S^{n-1})$  which is diffeomorphic to  $S^n$ . (The proof can be based on Munkres [13, §6.1].) Henceforth we will drop the prime, and identify  $\Gamma_n$  with  $\Gamma'_n$ .

The main properties of these groups can be described as follows.

*Theorem 4.* The group  $\Gamma_n$  is finite abelian for all  $n$ . For  $n \leq 6$  we have  $\Gamma_n = 0$ , but for  $n = 7$  the group  $\Gamma_7 \cong Z_{28}$  is non-zero.

In fact it follows from Theorem 3 that  $\Gamma_{4k-1} \neq 0$  for all  $k \geq 2$ .

The fact that  $\Gamma_n = 0$  for  $n \leq 3$  is due to Munkres. For  $n = 4$  this assertion has recently been announced by J. Cerf. In these cases the proof is based on the first definition, in terms of diffeomorphisms of  $S^{n-1}$ . For  $n \geq 5$  the proof is based rather on the second definition, in terms of twisted spheres (see [6]). Some indication of the method, for  $n \geq 5$ , is given in the following section.

#### 4. HOW TO RECOGNIZE AN HONEST SPHERE

Let  $M$  be a twisted  $n$ -sphere. How can we decide whether or not  $M$  is diffeomorphic to the standard  $n$ -sphere?

First choose an imbedding  $M \subset R^{n+k}$  for some large value of  $k$ . This is possible by a well-known theorem of Whitney [23].

*Lemma 5.* The normal bundle of  $M$  is trivial. That is, there exist  $k$  continuous linearly independent normal vector fields.

OUTLINE OF PROOF. Since  $M - x$  is contractible, it is certainly possible to choose such vector fields in the complement of  $x$ . Now the "obstruction" to extending over  $x$  is described by an element of the homotopy group  $\pi_{n-1}\text{SO}(k)$ . These groups (for  $k$  large) have been computed by Bott [1].

*Case 1.* If  $n \equiv 3, 5, 6$ , or  $7 \pmod{8}$ , then  $\pi_{n-1}\text{SO}(k) = 0$ , hence there is no obstruction.

*Case 2.* If  $n = 4i$  then the group  $\pi_{n-1}\text{SO}(k)$  is infinite cyclic. In this case the obstruction class can be identified with a multiple of the Pontrjagin number  $p_i[M]$ . (See [5].) But this number is zero by the signature theorem.

*Case 3.* If  $n \equiv 1$  or  $2 \pmod{8}$ , then  $\pi_{n-1}\text{SO}(k) \cong Z_2$ . The proof in this case is more delicate: the obstruction class  $\sigma$  satisfies  $J(\sigma) = 0$  according to [5], but J. F. Adams has shown that the  $J$  homomorphism has kernel zero for these values of  $n$ . This completes the proof.

Next we must introduce the concept of *framed cobordism*. By a *framing*  $\varphi$  of an imbedded manifold  $M^n \subset R^{n+k}$ ,  $k$  large, will be

meant a set of  $k$  linearly independent normal vector fields. Two framed manifolds  $(M_0, \varphi_0)$  and  $(M_1, \varphi_1)$  in  $R^{n+k}$  are called *framed cobordant* if there exists a compact framed manifold  $(W, \psi)$  of dimension  $n + 1$  in  $R^{n+k} \times [0, 1]$  with  $\partial W = M_0 \times 0 \cup M_1 \times 1$ , where the framing  $\psi$  restricts to  $\varphi_0$  at one end and to  $\varphi_1$  at the other. Using the disjoint sum as composition operation, the framed cobordism classes of  $n$ -manifolds form a group, which will be denoted by  $\Pi_n$ .

There are two fundamental theorems concerning these groups.

*Theorem of Pontrjagin [15].* The framed cobordism group  $\Pi_n$  is canonically isomorphic to the stable homotopy group  $\pi_{n+k}(S^k)$ ,  $n < k - 1$ .

*Theorem of Serre [18].* The stable homotopy groups  $\pi_{n+k}(S^k)$ ,  $0 < n < k - 1$ , are all finite abelian groups.

Thus  $\Pi_n$  is finite.

Consider the class of all framed twisted  $n$ -spheres in  $R^{n+k}$ . Using a suitably defined connected sum operation, and defining an appropriate concept of "isomorphism," we see that the set of all isomorphism classes of framed twisted  $n$ -spheres forms an abelian group, which will be denoted by  $\text{FG}_n$ . There is an exact sequence

$$\pi_n \text{SO}(k) \rightarrow \text{FG}_n \rightarrow \Gamma_n \rightarrow 0$$

PROOF. Every twisted sphere can be framed. The kernel of the homomorphism  $\text{FG}_n \rightarrow \Gamma_n$  is obtained by looking at standard spheres with exotic framings. But a framing of the standard  $S^n \subset R^{n+k}$  is clearly described by an element of  $\pi_n \text{SO}(k)$ .

On the other hand, there is clearly a homomorphism

$$j: \text{FG}_n \rightarrow \Pi_n \cong \pi_{n+k}(S^k)$$

Kervaire has shown that this homomorphism  $j$  is onto, except possibly when  $n \equiv 2 \pmod{4}$ . Thus every framed cobordism class contains a twisted sphere, except in dimensions 2, 6, 10, . . . .

Consider the kernel of this homomorphism  $j$ . Suppose, for example, that  $n \equiv 3 \pmod{4}$ . Then we claim that kernel  $(j)$  is infinite cyclic. Let  $(M, \varphi)$  be a framed twisted  $n$ -sphere belonging to the kernel of  $j$ . Then  $(M, \varphi) = \partial(W, \psi)$  where  $W \subset R^{n+k} \times [0, 1]$  is a compact framed manifold of dimension  $n + 1 = 4i$ . Hence the signature  $\sigma(W)$  is defined.

This integer  $\sigma(W)$  is an invariant of the framed manifold  $(M, \varphi)$ .

For if  $(M, \varphi)$  is also the boundary of  $(W', \psi')$ , then by placing  $W$  in  $R^{n+k} \times [0, 1)$  and  $W'$  in  $R^{n+k} \times (-1, 0]$ , we can construct a closed framed manifold

$$W \cup W' \subset R^{n+k} \times R$$

Giving  $W \cup W'$  the orientation which is compatible with that of  $W$ , we see that

$$\sigma(W \cup W') = \sigma(W) - \sigma(W')$$

But this signature must be zero: the fact that  $W \cup W'$  is framed implies that its Pontrjagin classes are all zero and hence, by the signature theorem, that its signature is zero.

Thus  $\sigma(W)$  is an invariant  $\bar{\sigma}$  of  $(M, \varphi)$ . In this way we define a homomorphism

$$\bar{\sigma}: \text{kernel}(j) \rightarrow Z$$

The construction of Section 2 shows that this homomorphism  $\bar{\sigma}$  is nontrivial.\*

Finally we claim that the kernel of  $\bar{\sigma}$  is zero. That is, if  $(M, \varphi)$  bounds  $(W, \psi)$  with  $\sigma(W) = 0$ , then  $M$  is diffeomorphic to  $S^n$ , and  $\varphi$  corresponds to the standard framing of  $S^n$ . The proof, which is fairly formidable, is divided into two parts. First, using the method of "surgery" [9], we show that all of the homotopy groups of  $W$  can be killed. In other words  $W$  can be replaced by a contractible manifold  $W'$ . Second, a theorem of Smale [19] asserts that such a manifold  $W'$  must be diffeomorphic to the  $(n+1)$ -disk. No further details are given here.

Thus  $\text{kernel}(j)$  is infinite cyclic. In other words there is an exact sequence

$$0 \rightarrow Z \rightarrow \text{FG}_{4i-1} \xrightarrow{j} \Pi_{4i-1} \rightarrow 0$$

It follows that  $\text{FG}_{4i-1}$  is an infinite abelian group of rank 1. But  $\pi_{4i-1}\text{SO}(k) \cong Z$  is also a group of rank 1. Using the sequence

$$\pi_{4i-1}\text{SO}(k) \rightarrow \text{FG}_{4i-1} \rightarrow \Gamma_{4i-1} \rightarrow 0$$

we finally see that the group  $\Gamma_{4i-1}$  is finite.

\* It is curious that  $\bar{\sigma}$  behaves somewhat differently in dimension 3 than in higher dimensions. Thus the image of  $\bar{\sigma}$  is generated by 8 if  $n = 4i - 1 \geq 7$ , but is generated by 16 if  $n = 3$ .

Similar constructions work for other values of  $n$ . Thus

$$0 \rightarrow F\Gamma_n \rightarrow \pi_{n+k}S^k \rightarrow 0 \quad \text{for } n \equiv 0 \pmod{4}$$

$$Z_2 \rightarrow F\Gamma_n \rightarrow \pi_{n+k}S^k \rightarrow 0 \quad \text{for } n \equiv 1 \pmod{4}$$

$$0 \rightarrow F\Gamma_n \rightarrow \pi_{n+k}S^k \rightarrow Z_2 \quad \text{for } n \equiv 2 \pmod{4}$$

$$0 \rightarrow Z_2 \rightarrow F\Gamma_n \rightarrow \pi_{n+k}S^k \rightarrow 0 \quad \text{for } n \equiv 3 \pmod{4}$$

For further information on these groups, the reader is referred to [6].

## REFERENCES

1. R. Bott, "The stable homotopy of the classical groups," *Ann. of Math.* **70** (1959) 313-337.
2. M. Brown, "The monotone union of open  $n$ -cells, is an open  $n$ -cell," *Proc. Amer. Math. Soc.* **12** (1961) 812-814.
3. J. Cerf, "Topologie de certains espaces de plongements," *Bull. Soc. Math. France* **89** (1961) 227-380.
4. F. Hirzebruch, *Neue topologische Methoden in der algebraischen Geometrie*, Berlin (1956) Springer.
5. M. Kervaire and J. Milnor, "Bernoulli numbers, homotopy groups and a theorem of Rohlin," *Proc. Intern. Congr. Math.*, Edinburgh (1958).
6. ———, "Groups of homotopy spheres: I," *Ann. of Math.* **77** (1963) 504-537.
7. J. Milnor, "On manifolds homeomorphic to the 7-sphere," *Ann. of Math.* **64** (1956) 399-405.
8. ———, "Sommes de variétés différentiables et structures différentiables des sphères," *Bull. Soc. Math. France* **87** (1959) 439-444.
9. ———, "A procedure for killing the homotopy groups of differentiable manifolds," *Symposia in Pure Math.* A.M.S. vol. III (1961) 39-55.
10. ———, "A survey of cobordism theory," *L'Enseignement Math.* **8** (1962) 16-23.
11. ———, *Morse theory*, Princeton, N.J. (1963) Princeton University Press.
12. M. Morse, "Relations between the numbers of critical points of a real function of  $n$  independent variables," *Trans. Amer. Math. Soc.* **27** (1925) 345-396.
13. J. Munkres, *Elementary differential topology*, Princeton, N.J. (1963) Princeton University Press.
14. R. Palais, "Extending diffeomorphisms," *Proc. Amer. Math. Soc.* **11** (1960) 274-277.
15. L. Pontrjagin, "Smooth manifolds and their applications in homotopy theory," *Trudy Mat. Inst. im. Steklov* **45** (1955), and A.M.S. translations, Series 2, **11**, 1-114.
16. G. Reeb, *Sur certaines propriétés topologiques des variétés feuilletées*, Paris (1952) Hermann.
17. R. Rosen, "A weak form of the star conjecture for manifolds," *Abstract* 570-28, *Notices Amer. Math. Soc.* **7** (1960) 380.

18. J.-P. Serre, "Homologie singulière des espaces fibrés. Applications," *Ann. of Math.* **54** (1951) 425-505.
19. S. Smale, "On the structure of manifolds," *Amer. J. Math.* **84** (1962) 387-399.
20. J. Stallings, "Polyhedral homotopy spheres," *Bull. Amer. Math. Soc.* **66** (1960) 485-488.
21. ———, "The piecewise linear structure of Euclidean space," *Proc. Cambridge Philos. Soc.* **58** (1962) 481-488.
22. H. Weyl, "Análisis situs combinatorio," *Revista Mat. Hispano-Americana*, **5-6** (1923-1924), 209-218ff.
23. H. Whitney, "Differentiable manifolds," *Ann. of Math.* **37** (1936) 645-680.
24. C. Zeeman, "The generalized Poincaré conjecture," *Bull. Amer. Math. Soc.* **67** (1961) 270.

# Lectures on Modern Mathematics

*In Three Volumes*

*Edited by T. L. SAATY, Arms Control and Disarmament Agency.*

All those desiring to keep abreast of the major achievements in the various fields of mathematics will welcome this work. Not only will it cover virtually every principle area of modern mathematics but it is written to reach as wide an audience as possible and to make readily accessible a maximum amount of material.

Each of the eminent contributors delineates a substantial research area. He describes it broadly and comprehensively and contributes to this description his individual evaluation of the aesthetic and practical aspects of the field, its position in mathematical development as a whole, and its future, as it is implied in the conjectural exposition of its unsolved problems.

The monographs are expanded versions of expository lectures, conceived and organized by the editor and given at the George Washington University under the sponsorship of the University and the Office of Naval Research. In writing the monographs, the lecturers have successfully compressed the enormous amount of material at their command and at the same time have preserved an informal approach, have refrained from any intricate excursions that might intimidate the novice, and have given each monograph the personal flavor of their own involvement in mathematical research.

## JOHN WILEY & SONS

NEW YORK • LONDON • SYDNEY

PRINTED  
IN  
U.S.A.

Saaty

Lectures on  
Modern Mathematics

II

510

LEC

Wiley